A Practical Transfer Learning Algorithm for Face Verification

Xudong Cao David Wipf Fang Wen Genquan Duan
{xudongca,davidwip,fangwen,genduan}@microsoft.com

Abstract

Face verification involves determining whether a pair of facial images belongs to the same or different subjects. This problem can prove to be quite challenging in many important applications where labeled training data is scarce, e.g., family album photo organization software. Herein we propose a principled transfer learning approach for merging plentiful source-domain data with limited samples from some target domain of interest to create a classifier that ideally performs nearly as well as if rich target-domain data were present. Based upon a surprisingly simple generative Bayesian model, our approach combines a KL-divergence-based regularizer/prior with a robust likelihood function leading to a scalable implementation via the EM algorithm. As justification for our design choices, we later use principles from convex analysis to recast our algorithm as an equivalent structured rank minimization problem leading to a number of interesting insights related to solution structure and feature-transform invariance. These insights help to both explain the effectiveness of our algorithm as well as elucidate a wide variety of related Bayesian approaches. Experimental testing with challenging datasets validate the utility of the proposed algorithm.

1. Introduction

Numerous computer vision applications involve testing a pair of facial images to determine whether or not they belong to the same subject. For example, this so-called *face verification* task is required by automatic PC or mobile logon using facial identity, or for grouping images of the same face for tagging purposes, etc. Important open algorithmic challenges include robustness across different platforms and environmental conditions, as well as computational efficiency for real-time implementation. Recently, several authors have demonstrated that simple, scalable generative Bayesian models are capable of achieving state-of-the-art performance on challenging face verification benchmarks [21, 18, 5]. The surprising success of these models (even surpassing much more complex discriminative approaches) is likely because facial appearances, when summarized by appropriate image descriptors or features, can be reasonably well-approximated by a linear summation of two independent latent factors: (i) intra-personal variations due to pose, expression and lighting, and (ii) variations due to differing identities. The former can be viewed as confounding, nuisance factors while the latter in isolation should determine successful face verification.

In [21, 18] these observations are exploited via a Bayesian factor analysis model called *probabilistic linear discriminant analysis* (PLDA), while the *Joint Bayesian* approach from [5] adopts a multivariate Gaussian distribution over image pairs achieving a similar effect. In fact, when trained with large-scale web images, the Joint Bayesian performance [5, 6] has even approached human capacity on the challenging LFW dataset [13]. For example, human performance on cropped faces is 97.53% [17]; the analogous performance achievable by the Joint Bayesian algorithm is 95.17% [6].

While these results are promising, many important practical scenarios involve cross-domain data drawn from potentially different facial appearance distributions. Therefore a model trained using widely available web images may suffer a large performance drop in an application-specific target domain that cannot be viewed as iid image samples from the web. The obvious solution would be to simply retrain with data from the relevant target domain; however, this often leads to over-fitting because available data are limited. This paper addresses these issues by deriving and analyzing a principled transfer learning algorithm for combining plentiful source-domain data (e.g., from the web, etc.) with relatively scarce target-domain samples. The underlying goal here is to match the idealistic performance achievable were a rich target-domain training set readily available.

Although conceptually we may address this problem by adapting any number of baseline face verification algorithms, we choose the Joint Bayesian algorithm as our starting point for two reasons. First, despite its simplicity and underlying Gaussian assumptions (see below for details), this algorithm nonetheless achieves the highest published results on the most influential benchmark face verification datasets. Secondly, the scalability and transparency of the Joint Bayesian cost function and update rules render principled transfer learning extensions and detailed analysis tractable.

Our basic strategy can be viewed from an information-theoretic perspective, where the idea is to penalize the Kullback-Leibler divergence between the distributions of source- and target-domain data to maximize the sharing of information. For the zero-mean multivariate Gaussians used by PLDA and Joint Bayesian algorithms, this reduces to the Burg matrix divergence between the corresponding covariance matrices [8]. This factor is then balanced with respect to the basic Joint Bayesian log-likelihood model which incorporates new samples from the target domain. The resulting model is optimized with respect to both factors using a highly efficient EM algorithm that easily scales to large problem sizes. Although the proposed model can be partially justified by the simplicity of the resulting update rules and conventional arguments for the utility of the KL divergence as a candidate regularizer, we produce a rigorous alternative rationalization by completely reformulating our model as a particular constrained rank minimization problem, leading to a variety of novel insights. To the best of our knowledge, Bayesian algorithms of this type have not be examined from such a perspective. The main contributions herein can then be summarized as follows:

- Development of a simple, scalable transfer learning method for adapting existing generative face verification models to new domains where data is scarce.
- Theoretical analysis connecting proposed model with a revealing, equivalent structured rank minimization problem. This demonstrates several desirable properties related to robustness, feature-transform invariance, subspace learning, and computational efficiency, while further elucidating many existing Bayesian face verification algorithms as a natural byproduct.
- Superior performance on challenging data sets representative of important applications of face verification.

The remainder of this paper is organized as follows. Section 2 describes related work on transfer learning while Section 3 briefly reviews the Joint Bayesian face verification algorithm which serves as the basis of our approach. The specifics of the proposed transfer learning algorithm are p-resented in Secion 4 followed by theoretical analysis and motivation for our particular model in Section 5. Finally, experimental results are carried out in Section 6.

2. Related Works

Transfer learning has been extensively studied in recent computer vision [22, 3, 23, 11, 16, 10].

Of particular relevance to our work, Kulis *et al.* learn a Mahalanobis distance function, where the learned metric is "close" to the Euclidian distance in the sense of Kullback-Leibler divergence [7]. This influential approach, termed ITML, has also been extended to domain adaptation problems [23, 16]. Alternatively, to resolve the domain difference for photo-sketch recognition, Xiaogang *et al.* proposed a coupled information-theoretic encoding method [28] to narrow the distribution gap of the encoded features. Other regularizers, including maximum mean discrepancy [20, 9] and Bregman divergence [24], have also been studied for transfer learning. Our algorithm differs from these discriminative approaches via the choice of our generative model and its subsequent interaction with the KL divergence regularizer.

Transfer learning algorithms have also been developed based upon recent rank minimization techniques [4, 14, 15]. However, these methods apply to problems that are structurally very different from face verification and existing methods do not apply here. Although our algorithm is not directly derived from a rank minimization perspective, as intimated above it can be interpreted as a particular minimization task that includes multiple concave penalties on matrix singular values that are combined in a novel way.

3. Review of the Joint Bayesian Method

This section briefly reviews the Joint Bayesian method for face verification [5] which will serve as the basis for our transfer learning algorithm. In this context we assume that the appearance of relevant facial features is influenced by two latent factors: identity and intra-personal variations. This fact is commonly approximated by

$$x = \mu + \epsilon, \tag{1}$$

where x is the assumed facial feature, e.g., LBP, SIFT, etc., which is linearly decomposed into two independent variables related to identity μ and intra-personal variations ϵ. The Joint Bayesian method then models both μ and ϵ as multivariate Gaussians with zero mean (after the appropriate centering operation) and covariance matrices S_μ and S_ϵ respectively. S_μ and S_ϵ can be interpreted as the between-class and within-class covariances, and both can be computed empirically via sample averages given sufficient data, or estimated more accurately by an EM algorithm [5]. During the testing phase, unlike previous Bayesian face recognition algorithms which discriminate based on the difference between a pair of faces [19, 27], the Joint Bayesian classifier is based upon the full joint distribution of face image pairs leading to a considerable performance boost.

Specifically, let H_I represent the intra-personal hypothesis (same identity) and H_E represent the extra-personal hypothesis (different identity). Using (1), it is readily shown that the joint distributions $p(x_1, x_2|H_I)$ and $p(x_1, x_2|H_E)$

are zero-mean Gaussians with covariance matrices

$$\begin{bmatrix} S_\mu + S_\varepsilon & S_\mu \\ S_\mu & S_\mu + S_\varepsilon \end{bmatrix} \text{ and } \begin{bmatrix} S_\mu + S_\varepsilon & 0 \\ 0 & S_\mu + S_\varepsilon \end{bmatrix}$$

respectively. Given these distributions, the likelihood ratio test

$$r\left(x_1, x_2\right) = \log \frac{P\left(x_1, x_2 | H_I\right)}{P\left(x_1, x_2 | H_E\right)} \qquad (2)$$

represents a natural classification criterion. The resulting decision function can be reduced to a convenient closed form leading to an efficient algorithm for the testing phase.

4. Transfer Learning Algorithm

We will now adapt the Joint Bayesian model to the transfer learning problem by first proposing an appropriate cost function followed by the development of a simple EM algorithm for training purposes.

4.1. Information-Theoretic Cost Function

Given the basic Joint Bayesian model with parameters $\Theta_s = \{S_\mu, S_\epsilon\}$ fitted to source-domain data, and a handful of labeled target-domain data $\mathcal{X}$, the underlying goal is to learn a new model with analogous parameters $\Theta_t = \{T_\mu, T_\epsilon\}$ that adequately reflects both domains, and in particular, generalizes to new target-domain data. In the absence of source-domain data, the unknown parameter Θ_t could be estimated by optimizing $\log p(\mathcal{X}|\Theta_t)$ over $\mathcal{X}$, where the likelihood model for $\mathcal{X}$ is simply analogous to the Joint Bayesian one. Of course when the available training samples are limited, over-fitting is likely and generalization performance on unseen data will be poor. When additional source-domain data are accessible, however, we may ameliorate the risk of over-fitting by including an additional regularizer, or prior, that penalizes deviations from the distribution of source data. From an information-theoretic perspective, the KL divergence, which quantifies the information lost when we approximate the target-domain distribution with the source-domain distribution, represents a useful candidate regularizer for this task. After combining with the log-likelihood term, this results in the optimization problem

$$\min_{\Theta_t} \ -\log p(\mathcal{X}|\Theta_t) + \lambda \, \mathrm{KL}(p(\mathcal{X}|\Theta_t)||p(\mathcal{X}|\Theta_s)), \qquad (3)$$

where the parameter λ balances the relative importance between the new observations and the prior knowledge.

The KL divergence, as well as alternative penalties based on Bregman divergences and maximum mean discrepancy, have been motivated for related transfer learning purposes [7, 23, 16, 20, 24, 9], although not in combination with a likelihood function as we have done here. However, the primary advantage of (3) in particular is threefold:

1. In the absence of significant source-domain data, (3) reduces to a current state-of-the-art algorithm for face verification.
2. Based on an EM framework, the KL divergence, when coupled with the Joint Bayesian distributional assumptions, leads to simple closed-form update rules that scale to large-sized problems. Moreover, the learned model is suitable for real-time applications at test time.
3. A completely independent justification of (3) and the associated EM update rules is possible using ideas from convex analysis (see Section 5).

4.2. Optimization via EM Updates

Assuming the samples from different target-domain subjects are independent, the joint distribution of all samples can be factorized, such that (3) reduces to

$$\min_{\Theta_t} \ -\sum_i \log p(X_i|\Theta_t) + \lambda \sum_i \mathrm{KL}(p(X_i|\Theta_t)||p(X_i|\Theta_s)), \qquad (4)$$

where X_i is a long column vector formed by concatenating all of the samples from subject i. $p(X_i|\Theta)$ then represents the corresponding likelihood function from the Joint Bayesian model, which involves a potentially huge, zero-mean Gaussian with covariance $P_i \Omega_i P_i^T$, where $\Omega_i = \mathrm{diag}[T_\mu, T_\varepsilon, \cdots, T_\varepsilon]$ is block diagonal with T_ε repeated once for each sample from subject i, and

$$P_i = \begin{bmatrix} \mathbf{I} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}.$$

Directly minimizing (4) is difficult because of the high-dimensionality, the coupling between the log-likelihood and KL terms, and the fact that the unknown parameters must be positive semi-definite, symmetric matrices. To address these issues, we present a generalized EM algorithm which simplifies the objective function by introducing additional latent variables based on (1) that allow us to conditionally factorize the distribution of all samples for each subject.

Specifically, we decompose all of the samples of one subject[1] X into two latent parts based on (1): identity μ, which is invariant for all images of the same subject, and intra-personal variations $\{\epsilon_1, ...\epsilon_m\}$ assuming m images of this subject. These latent variables can be expressed as a long column vector $H = \{\mu; \epsilon_1; ...; \epsilon_m\}$, where $X = PH$. As the identity and intra-personal variations are independent Gaussians, it is easy to show that H follows a zero-mean Gaussian distribution with covariance Ω. We may

[1] As the discussion in the remainder of this section is based on a single subject, the subject index has been omitted.

now optimize the objective function in (3) by iteratively computing the expectation of the latent variables (E step) and updating the parameters by maximizing the expected penalized log-likelihood found in the E step (M step). Due to limited space, we omit the derivation and directly present the closed-form solutions in for the E and M steps. Readers can refer to the supplementary file for the derivations.

E-step: Given the samples of one subject X, the expectation of the associated latent variable H can be derived as

$$E(H|X) = \Omega P^T (P\Omega P^T)^{-1} X. \tag{5}$$

Note that although the required multiplications and inversions involve high-dimenstional matrices, the problem structure can be exploited such that only $O(d^3 + md^2)$ computations are required, where d is the feature dimension of each image [5].[2] More importantly however, the overall complexity required to evaluate (2) for testing new image pairs is at most $O(\text{rank}[T_\mu]d) \leq O(d^2)$ (see Section 5).

M-step: $\Theta_t = \{T_\mu, T_\epsilon\}$ is updated via

$$\begin{aligned} T_\mu &= w\, S_\mu + (1-w)n^{-1} \sum_i \mu_i \mu_i^T \\ T_\epsilon &= w\, S_\epsilon + (1-w)k^{-1} \sum_j \epsilon_j \epsilon_j^T, \end{aligned} \tag{6}$$

where $w = \lambda/(1+\lambda)$, n is the number of subjects, and k is the total number of images of all subjects. These intuitive updates reveal that source- and target-domain information are merged by a weighted linear combination. Also, from an implementational standpoint, instead of adapting the mean face from the source to target domain, we directly estimate the mean using only target-domain data. This is because first-order statistics can be reliably estimated with relatively limited data, even though the second-order, high-dimensional covariances cannot be. Finally, we should mention that the empirically observed convergence rate is very fast, e.g., it only takes around five iterations to produce all of the experiments in Section 6.

5. Low-Rank Interpretation and Analysis

Let m_i denote the number of images of subject i such that $k = \sum_{i=1}^n m_i$. Then we use $\mathbb{X}$ to denote the $d \times k$ matrix of all image features from all subjects, with the j-th column x_j representing the feature vector of image j. Also, define Ψ to be the $n \times k$ matrix with i-th row given by all zeros except a vector of m_i ones starting at element index $e_i = \sum_{r=1}^{i-1} m_r + 1$.

We now present a novel reformulation of the proposed learning algorithm that we later show provides insights into the types of solutions that will be favored. For space considerations the proof, which is based upon ideas from convex analysis, has been deferred to the supplementary material.

Theorem 1 *The iterations from (5) and (6) are guaranteed to reduce (or leave unchanged once a stationary point is reached) the minimization problem*

$$\begin{aligned} \min_{\mathbb{E},\mathbb{M}} \quad & n \log |T_\mu| + k \log |T_\epsilon| \\ \textit{s.t.} \quad & \mathbb{X} = \mathbb{E} + \mathbb{M}\Psi, \\ & T_\mu = \tfrac{1}{n}\mathbb{M}\mathbb{M}^T + \lambda S_\mu, \quad T_\epsilon = \tfrac{1}{k}\mathbb{E}\mathbb{E}^T + \lambda S_\epsilon. \end{aligned} \tag{7}$$

The remainder of this Section will argue that the optimization problem from (7) provides a compelling, complementary picture of the original transfer learning formulation from Section 4. As a natural byproduct, it also elucidates the behavior a number of related Bayesian algorithms including PLDA [21, 18] and Joint Bayes [5].

To begin, the penalty terms in (7) both rely on the log-det function, which represents a somewhat common surrogate for the matrix rank function. This relationship can be understood as follows. For a given symmetric, positive semi-definite matrix Z, let $\boldsymbol{\sigma}$ denote the vector of all singular values in Z (which will be non-negative) and σ_r its r-th element. We then have

$$\log|Z| = \tag{8}$$

$$\sum_r \log \sigma_r = \lim_{p \to 0} \frac{1}{p} \sum_r (\sigma_r^p - 1) \propto \|\boldsymbol{\sigma}\|_0 = \text{rank}[Z],$$

In this context, $\log|Z|$ can be viewed as a scaled and translated version of rank$[Z]$.

Now for simplicity, first assume that no prior source-domain knowledge is available, and thus $S_\mu = S_\epsilon = 0$. The objective function from (7) is basically attempting to find covariances T_μ and T_ϵ of (approximately) minimal rank, subject to the constraint that the latent variables $\mathbb{M}$ and $\mathbb{E}$, when confined to the subspaces proscribed by their respective covariances, satisfy the constraint $\mathbb{X} = \mathbb{E} + \mathbb{M}\Psi$. Here columns of $\mathbb{E}$ and $\mathbb{M}$ each correspond with intra-personal and extra-personal variation components respectively.

Low rank solutions can be highly desirable for regularization purposes, interpretability, and implementational efficiency. The latter is especially crucial for many practical applications, where minimal rank implies fast evaluation on test data (see below). However, in the absence of

[2] Technically speaking, this algorithm only computes an approximate E-step, and hence falls into the wider category of generalized EM and MAP EM algorithms [12]. While the full E-step can actually be calculated using our model with limited additional computation (we merely need to compute a posterior covariance analogous to the mean from (5)), we choose not to include this extra term for several reasons. First, generalized EM algorithms enjoy similar convergence properties to regular EM and are widely used in machine learning. Second, we have observed empirically that the performance is essentially unchanged with or without this additional covariance term. And finally, omitting this covariance leads to much more transparent analysis (see Section 5). The supplementary file contains more information about these distinctions.

prior knowledge, and with limited training data, the associated subspace estimates may be unreliable or possibly associated with undesirable degenerate solutions. Fortunately, when prior information is available in the form of nonzero covariances S_μ and S_ϵ, the situation becomes much more appealing. The log-det penalty now handles the subspace spanned by the prior source-domain information (meaning the span of the singular vectors of λS_μ and λS_ϵ that have significant singular values) very differently than the orthogonal complement. In directions characterized by small (or zero) singular values, T_μ or T_ϵ will be penalized heavily akin to the rank function per the analysis above. In contrast, when source-domain singular values are relatively large, the associated penalty softens considerably, approaching a nearly-flat convex, ℓ_1 norm-like regularizer (in the sense that $\log(\sigma + c)$ achieves a near constant gradient with respect to σ as c becomes large).

Thus to summarize then, the source-domain information essentially encodes the relative curvature, or severity, of the penalization on T_μ or T_ϵ: strong, non-convex penalization of directions orthogonal to the significant singular vectors of λS_μ and λS_ϵ, while relaxing to very mild, nearly convex penalty elsewhere. This all implies relative freedom to explore regions supported by prior information, but a stricter impediment for drifting into novel regions without sufficient data to support it. Other benefits are as follows.

Invariance: The reformulation (7) highlights the fact that the proposed approach is invariant to invertible linear transformations of the feature vectors in the following sense: If T_μ^* and T_ϵ^* are computed from the optimal solution to (7), then WT_μ^* and WT_ϵ^* are the optimal solution when $x \to Wx$, with W invertible. This occurs as a direct consequence of the fact that $\log|AB| = \log|A| + \log|B|$ for two matrices A and B, allowing a simple reparameterization to reveal the stated invariance. Moreover, it is readily shown that, because the likelihood ratio test (2) used to compare two new faces is invariant to an invertible transformation such as W, the solution WT_μ^* and WT_ϵ^* is for all practical purposes fully equivalent to T_μ^* and T_ϵ^*. This highly desirable invariance property is quite unlike other sparse or low-rank models that incorporate, for example, convex penalties such as the ℓ_1 norm or the nuclear norm. With these penalties an invertible feature transform would lead to an entirely different decision rule and therefore different classification results.

Efficiency: It is straightforward to show that the test statistic (2) can be compactly represented as $x_1^T A x_1 + x_2^T A x_2 + x_1^T B x_2$ where A and B are matrices that depend only on T_ϵ and T_μ. Importantly, it can be shown that

$$\text{rank}[A] \le \text{rank}[T_\mu] \quad \text{and} \quad \text{rank}[B] \le \text{rank}[T_\mu]. \qquad (9)$$

The derivation of these bounds is contained in the supplementary file. Consequently, (9) implies that, provided at least T_μ is of low rank, then (2) can be computed efficiently using convenient low-rank formula. This is crucial for any number of practical applications, e.g., face log-on for smartphones, where fast, real-time computations are required. We note that the convex nuclear norm substitution for the rank penalty does not shrink nearly as many singular values to exactly zero (experiments not shown), and thus does not produce nearly as parsimonious a representation. Thus heuristic singular value thresholding is required for a practical, computationally-efficient implementation.[3]

6. Experiments

Here we present empirical results that validate our algorithm and elucidate its desirable properties.

Source-Domain Model: Assembled from Internet photos, the large-scale WDRef dataset [5] is characterized by large variations in pose, expression, and lighting and therefore reflects the diversity of intra- and extra-personal variations (see Figure 1 for examples). It contains 99,773 images of 2,995 subjects, far more than even LFW. We use WDRef (provided by the authors) as the source-domain dataset and adopt the method in [5] to learn the source-domain model used in all experiments.

High-Dim Feature: We use the recently proposed high-dimensional LBP [6] as the feature representation x. It densely samples multi-scale LBP descriptors centered at dense facial landmarks, and then concatenates them to form a high-dimensional feature. Due to its high dimensionality (over 100,000), PCA is applied to reduce the size to a feasible range for subsequent learning.

Experimental Protocol: For all experiments the target-domain data is divided into three parts for training, validation and testing. The validation set is used for selecting the source/target-domain trade-off parameter λ as well as the threshold applied to the likelihood ratio test from (2). The separate testing set is used only for generating the final results presented in this section. Typically, there is no identity overlap among the three parts, unless otherwise stated.

6.1. Results with Similar Source/Target Domains

A good transfer learning method should seamlessly perform well even with differing degrees of similarity between the source and target domains. This section explores the case where there exists substantial similarity between the two domains, while Section 6.2 will examine the alternative. The PCA dimension is fixed to 2,000 for all of these experiments.

[3] For practical purposes, and to avoid undefined solutions involving the log of zero in (7), both S_ϵ and S_μ can be chosen with no strictly zero-valued singular values. This would then imply that T_ϵ and T_μ, and therefore A and B cannot be strictly low rank without some minimal level of thresholding. However, this is a relatively minor implementational detail and does not affect the overall nature of the arguments made in this section.

Figure 1. Typical samples of the datasets used in our experiments. From left to right: WDRef, LFW, Video Camera Dataset, Family Album Dataset, and WDAsian.

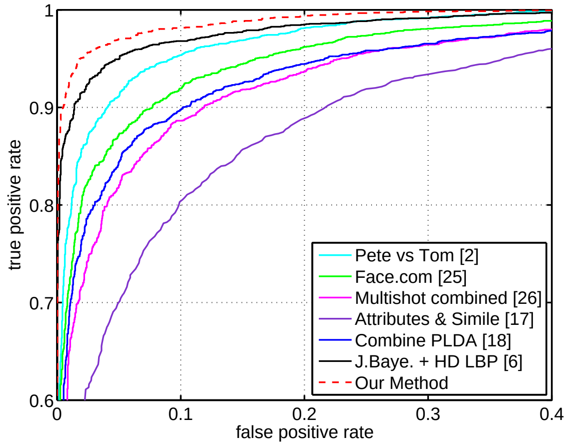

Figure 2. The ROC Curve on LFW. Our approach exceeds the best published result under the unrestricted protocol.

We use the LFW dataset [13] for the target domain both because of its similarity with WDRef and because it represents a well-studied and challenging benchmark allowing us to place our results in the context of existing face verification systems [17, 26, 25, 2, 18, 6]. Figure 2 shows the results of our method compared with state-of-the-art algorithms. The previous best published result on the LFW data (95.17%, unrestricted protocol) is achieved by the Joint Bayesian method [5] trained using WDRef and the high-dim features from [6]. Although WDRef and LFW data are similar, this result shows that the proposed transfer learning method can still improve the accuracy even further to **96.33%**, which is noteworthy given that human performance is 97.53% [17]. Moreover, given that our algorithm explicitly abides by all of the rules pertaining to the unrestricted LFW protocol, it now represents the best reported result on this important benchmark. We also emphasize that the top performing methods on LFW all use outside training data as we have; however, our algorithm appears to be the most effective at assimilating discriminative information.

6.2. Results with Large Domain Differences

Next we experimentally verify the proposed method in two common daily-life scenarios where, unlike the previous section, considerable domain differences exists relative to source-domain data collected from the Internet.

Video camera dataset. The images in this dataset were collected by video camera under various challenging conditions. It contains 58 subjects and 1,948 images in total, typically around 40 samples per subject. Figure 1 displays some typical exemplars. We study the accuracy of our model as a function of the number of subjects used in the target domain. Due to significant domain differences between web images and the captured video camera frames, the error rate of the baseline source-domain model is 13.3%. However, with data from only 16 subjects, our method can effectively reduce the error to 5.8% as shown in Table 1.

♯ Subj.	2	4	8	16
TDO	18.5%	15.3%	12.2%	10.2%
TL	9.8%	8.5%	7.0%	5.8%

Table 1. Error rates using video camera data. Results from models trained with target-domain data only (TDO) and transfer learning (TL). The source-domain model error rate is 13.3%.

Family Album dataset. This dataset contains eight real family photo albums collected from personal contacts. There is considerable diversity between the different albums in terms of the number of images, subjects, and time frame. The smallest album contains 10 subjects and around 400 images taken over the past two years. In contrast, the largest albums contain hundreds of subjects and around $10,000$ images taken over the past eight years. To mimic a practical scenario, we consider each family album as a target domain. In each case, the images taken earlier (20% of the whole album) are used for training. Unlike the settings in other experiments, the identities are overlapped in training and testing sets. As shown in Figure 3, for most albums the error rate is reduced to less than half of the error rate achieved by the source-domain baseline model. We expect that this could improve the user experience in personal album management on many platforms such as PC, phone, and social networks.

6.3. Comparisons with Existing TL methods

Using the video camera dataset from the previous section, we now turn to comparisons with competing transfer

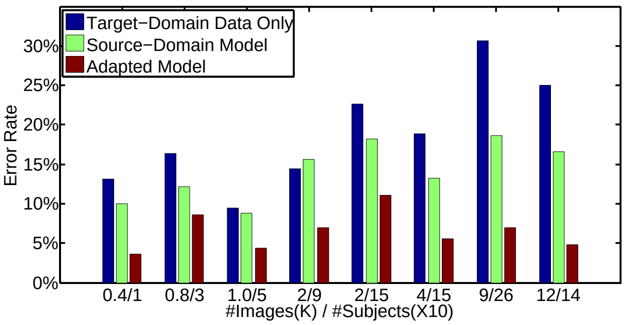

Figure 3. Error rates using family albums. X-axis labels show the number of images and subjects in the corresponding album.

learning techniques. Because metric and subspace learning represent influential, extensible approaches for face verification, we choose to conduct experiments with two popular representatives. First, information-theoretic metric learning (ITML) [7] learns a Mahalonobis distance that is "close" to a prior metric. By using the source-domain to obtain such a prior, ITML is naturedly extended as a transfer learning method, referred to as *T-ITML*. Secondly, linear discriminative analysis (LDA) [1] learns a discriminative subspace for differentiating different identities. Recently Si *et al.* [24] proposed a framework for transductive transfer subspace learning based on Bregman divergences. By applying their framework, we then have transfer LDA, or *T-LDA*, as a useful competing method.

We use 20 subjects for transfer learning. For *T-ITML*, a large amount training pairs[4] are generated to maximize its performance, while for *T-LDA* the optimal subspace dimensionality must be selected. All other parameters are set according to published recommendations [7, 24]. As shown in Table 2, our method leads to the lowest error rate. This occurs because our improvement over the baseline source-domain model is large (around 10%), and because our source model is stronger to begin with. Additionally, it should be mentioned that *T-ITML* involves a computationally intensive training procedure because of the high number of training pairs required.

PCA Dim.	T-LDA	T-ITML	Ours
100	17.5(20.9)	12.1(21.5)	**8.2**(18.3)
200	16.8(19.1)	11.4(19.8)	**6.2**(16.9)
400	15.4(18.2)	10.8(18.7)	**5.4**(15.1)

Table 2. Error rates of different transfer learning methods. The result in brackets is obtained by the corresponding source model.

6.4. Transfer Learning vs. Large Scale Data

When provided with a small amount of target-domain data, ideally a good transfer learning algorithm will produce a new model which performs nearly as well as a model trained using a fully-representative, large-scale target-domain dataset. To assess this issue, we experimentally study the performance of our proposed model as a function of the amount of the target-domain data. We construct a large-scale dataset named WDAsian as the target domain. It contains 2,000 *Asian* celebrities and around 100,000 images in total. Over 1,800 subjects have 40+ samples. Figure 1 shows some typical samples.

Two conclusions can be drawn from the results in Table 3. First, our transfer learning model performs similarly to the model trained using the large-scale target-domain data, which is around *20 times* larger than the data used for transfer learning. Secondly, when the target-domain data is scarce, the over-fitting problem is so severe that the model trained only with target-domain data performs much worse than the source-domain model. Actually, this conclusion holds generally as discussed further in Section 6.5.

♯ Subj.	20	80	640	1280	1800
TDO	33.0%	28.8%	13.0%	11.2%	9.5%
TL	11.5%	11.1%	9.9%	9.5%	9.1%

Table 3. The error rate of the models trained with target-domain data only (TDO) and transfer learning (TL). The error rate of source-domain model is 13.6%.

6.5. The Role of Regularization

The amount of target-domain data and model complexity jointly determine the risk of over-fitting and the need for regularization. To examine these effects we use the PCA dimension to represent the freedom of the transfer learning model, while WDAsian is used as the target-domain dataset. The source-domain weight is fixed to 0.8.

High model complexity generally translates into high structural risk and over-fitting. Figure 4 (*left*) clearly demonstrate this. When the data is scarce, the model with higher PCA dimension performs worse when trained on target-domain data alone. With increasing data however the trend reverses; the model with higher dimension eventually outperforms the others because the structural risk is no longer the dominating factor, i.e., the more complex model can harness the extra information to improve performance. Our method presents another way of hedging the structural risk. As shown in Figure 4 (*right*), even when training with small amounts of target-domain data (only 10 subjects), the higher dimensional model performs best with limited effects of over-fitting. The reason of course is that the source-domain data acts as a powerful regularizer, centering the solution space at a more reasonable baseline. In this con-

[4] We generate 200,000 pairs for training the source-domain metric and 6,000 pairs for transfer learning

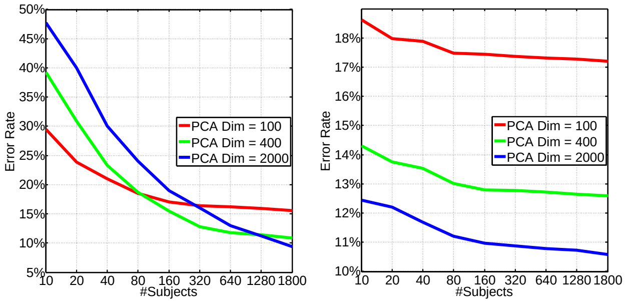

Figure 4. *Left*: the results obtained with target-domain data only. *Right*: the results obtained by transfer learning.

text the higher PCA dimensions can be safely exploited to enhance the performance without over-fitting.

7. Conclusion

This paper presents a generative Bayesian transfer learning algorithm particularly well-suited for the face verification problem. This is possible in large part because large web-based facial databases contain a variety of relevant information that can be used to bias estimation in smaller, more nuanced, application-specific domains. Although admittedly quite simple, our extensive theoretical and empirical analysis suggest that it nonetheless represents a viable candidate for many practical, real-time systems.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *PAMI*, 1997. 7

[2] T. Berg and P. N. Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *BMVC*, 2012. 6

[3] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. *NIPS*, 2010. 2

[4] B. Chen, W. Lam, I. W. Tsang, and T.-L. Wong. Discovering low-rank shared concept space for adapting text mining models. *PAMI*, 2013. 2

[5] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: a joint formulation. In *ECCV*, 2012. 1, 2, 4, 5, 6

[6] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, 2013. 1, 5, 6

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007. 2, 3, 7

[8] J. Dhillon. Differential entropic clustering of multivariate gaussians. In *NIPS*, 2007. 2

[9] B. Geng, D. Tao, and C. Xu. Daml: Domain adaptation metric learning. *Image Processing, IEEE Transactions on*, 20(10):2980–2989, 2011. 2, 3

[10] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 2

[11] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011. 2

[12] R. J. Hathaway. Another interpretation of the em algorithm for mixture distributions. *Statistics & Probability Letters*, 4(2):53–56, 1986. 4

[13] G. Huang, M. Mattar, T. Berg, E. Learned-Miller, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in Real-Life Images: Detection, Alignment, and Recognition*, 2008. 1, 6

[14] B. Hutchinson, M. Ostendorf, and M. Fazel. Low rank language models for small training sets. *Signal Processing Letters, IEEE*, 18(9):489–492, 2011. 2

[15] I.-H. Jhuo, D. Liu, D. Lee, and S.-F. Chang. Robust visual domain adaptation with low-rank reconstruction. In *CVPR*, 2012. 2

[16] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011. 2, 3

[17] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1, 6

[18] P. Li, Y. Fu, U. Mohammed, J. Elder, and S. Prince. Probabilistic models for inference about identity. *PAMI*, 34(1):144 –157, 2012. 1, 4, 6

[19] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000. 2

[20] S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *Proceedings of the 23rd national conference on Artificial intelligence*, volume 2, pages 677–682, 2008. 2, 3

[21] S. J. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *ICCV*, 2007. 1, 4

[22] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. In *CVPR*, 2008. 2

[23] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. 2010. 2, 3

[24] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(7):929–942, 2010. 2, 3, 7

[25] Y. Taigman and L. Wolf. Leveraging billions of faces to overcome performance barriers in unconstrained face recognition. arXiv:1108.1122, 2011. 6

[26] Y. Taigman, L. Wolf, and T. Hassner. Multiple one-shots for utilizing class label information. In *BMVC*, 2009. 6

[27] X. Wang and X. Tang. A unified framework for subspace face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1222–1228, 2004. 2

[28] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, 2011. 2