Supplementary Materials for "A Practical Transfer Learning Algorithm for Face Verification"

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

1. Derivation of EM Optimization

Here we briefly describe more details of the proposed EM algorithm.

E-step: As discussed in the main text, H_i and X_i represent the hidden and observed variables respectively associated with subject i, where we have reintroduced subject indeces to clarify the derivation. As is conventional, the E-step involves computing the expected value of the complete log-likelihood over the posterior distribution $p(H_i|X_i, \Theta_t)$ for all i, meaning

$$-\sum_i E_{p(H_i|X_i,\Theta_t)} \log\ p(X_i, H_i|\Theta_t) \equiv -\sum_i E_{p(H_i|X_i,\Theta_t)} \log\ p(H_i|\Theta_t), \tag{1}$$

where the equivalency holds because $p(X_i|H_i)$ is independent of Θ_t, i.e., $X_i = P_i H_i$ based on Section 4 of our submission. Here $p(H_i|\Theta_t)$ is a zero-mean Gaussian, with block-wise diagonal covariance matrix $\Omega_i = \text{diag}[T_\mu, T_\epsilon, \cdots, T_\epsilon]$, where T_ϵ is repeated once for each sample from subject i. Note that because the KL divergence term is independent of the hidden variables, we may ignore it for the E-step. Also, we will henceforth assume that all expectations are with respect to the appropriate posterior $p(H_i|X_i, \Theta_t)$ and omit further explicit reference.

Using basic properties of expectations, and the Gaussian structure of $p(H_i|\Theta_t)$, we have

$$-\sum_i E[\log\ p(H_i|\Theta_t)] \equiv -\sum_i \left[\text{trace}(\Omega_i^{-1} E[H_i H_i^T]) - \log|\Omega_i|\right], \tag{2}$$

where we have omitted irrelevant factors independent of H_i or Θ_t. So the E-step reduces to simply computing the expected value of $H_i H_i^T$ for all i. In general, we know that $E[H_i H_i^T] = Cov[H_i] + E[H_i]E[H_i]^T$; however, we adopt the simplifying assumption $E[H_i H_i^T] \approx E[H_i]E[H_i]^T$, where

$$E[H_i] = \Omega P_i^T (P_i \Omega_i P_i^T)^{-1} X_i. \tag{3}$$

This expression is obtained by computing the conditional mean of $p(H_i|X_i, \Theta_t)$, which is available via standard formula given the parameters of the joint distribution $p(H_i, X_i|\Theta_t) = p(X_i|H_i)p(H_i|\Theta_t)$.

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 (3)), we choose not to include this extra term for several reasons. First, generalized EM algorithms (of which our approximation is a special case) enjoy similar convergence properties to regular EM and are widely used in machine learning. Secondly, we have observed empirically that the performance is essentially unchanged with or without this additional covariance factor, largely because this factor tends to be very small in practice. And finally, removing this covariance leads to much more transparent analysis. Regardless, in Section 4 below we will closely examine the theoretical ramifications of this approximation.

M-step: Ignoring irrelevant constant terms, the KL divergence term can be expanded as

$$\text{KL}(p(H_i|\Theta_t)||p(H_i|\Theta_s)) = \text{trace}\left[S_\mu T_\mu^{-1} + m_i S_\epsilon T_\epsilon^{-1}\right] - \log\left|S_\mu T_\mu^{-1}\right| - m_i \log\left|S_\epsilon T_\epsilon^{-1}\right|, \tag{4}$$

where m_i is the number of images belonging to subject i. The M-step then involves minimization of the sum of (2) and (4) over $\Theta_t = \{T_\mu, T_\epsilon\}$. After a series of algebraic manipulations, we arrive at the optimal values

$$\begin{aligned} T_\mu &= wS_\mu + (1-w)\sum_i E[\mu_i]E[\mu_i]/n \\ T_\epsilon &= wS_\epsilon + (1-w)\sum_i\sum_j E[\epsilon_{ij}]E[\epsilon_{ij}]/\sum_i m_i, \end{aligned} \tag{5}$$

where $w = \lambda/(1+\lambda)$, n is the number of subjects, and the expectations are obtained from the E-step.

2. Proof of Theorem 1

The proof is based on the application of basic principles from convex analysis. Consider the alternative optimization problem motivated by Fenchel duality

$$\begin{aligned} \min_{\mathbb{M},\mathbb{E};\ \Psi,\Gamma\succeq 0} \quad & \text{trace}\left[\left(\mathbb{M}\mathbb{M}^T + n\lambda S_\mu\right)\Psi^{-1}\right] + \text{trace}\left[\left(\mathbb{E}\mathbb{E}^T + k\lambda S_\epsilon\right)\Gamma^{-1}\right] + \\ & k\left|\Gamma\right| + n\log\left|\Psi\right| \\ \text{s.t.} \quad & \mathbb{X} = \mathbb{E} + \mathbb{M}\Phi, \end{aligned} \tag{6}$$

which can be iteratively minimized by coordinate descent over $\mathbb{M}$, $\mathbb{E}$, Ψ, and Γ. First consider minimization over Ψ and Γ with $\mathbb{M}$ and $\mathbb{E}$ fixed. Convenient, closed-form solutions are available by taking the gradient of the objective function and setting it to zero. This produces the updates

$$\begin{aligned} \Psi &\leftarrow \frac{1}{n}\mathbb{M}\mathbb{M}^T + \lambda S_\mu \\ \Gamma &\leftarrow \frac{1}{k}\mathbb{E}\mathbb{E}^T + \lambda S_\epsilon. \end{aligned} \tag{7}$$

Now we optimize over $\mathbb{M}$ and $\mathbb{E}$ with Ψ and Γ fixed, producing the quadratic problem

$$\begin{aligned} \min_{\mathbb{M},\mathbb{E}} \quad & \text{trace}\left[\mathbb{M}^T\Psi^{-1}\mathbb{M} + \mathbb{E}^T\Gamma^{-1}\mathbb{E}\right] \\ \text{s.t.} \quad & \mathbb{X} = \mathbb{E} + \mathbb{M}\Phi, \end{aligned} \tag{8}$$

Here the problem decouples across each subject i giving

$$\min_{H_i} H_i^T\Omega_i^{-1}H_i \quad \text{s.t. } X_i = P_iH_i, \tag{9}$$

where we have borrowed the definitions of X_i, P_i, and H_i from our original submission (with an allowance for differing numbers of images within each subject, and hence the subject index i must be reintroduced). Also, we define $\Omega_i = \text{diag}[\Psi, \Gamma, \ldots, \Gamma]$, where Γ is replicated m_i (the number of images for subject i) times. From basic linear algebra we know that the minimizing solution then becomes

$$H_i \leftarrow \Omega_i P_i^T\left(P_i\Omega_i P_i^T\right)^{-1}X_i, \tag{10}$$

from which the optimal $\mathbb{M}$ and $\mathbb{E}$ can be constructed by combining all H_i. The updates from (7) and (10) are equivalent to the EM updates from Section 4 in the main text up to an irrelevant scale factor of $(1-w)$ missing in (7) that can always be canceled out by a simple reparameterization.

Thus, our EM algorithm is guaranteed to iteratively minimize (6), and it only remains to show the relationship between (6) and the optimization problem presented in the theorem statement from the main text. This is straightforward when we plug the optimal values of Ψ and Γ in to (6) leading to the optimization problem.

$$\begin{aligned} \min_{\mathbb{M},\mathbb{E}} \quad & n\log\left|\tfrac{1}{n}\mathbb{M}\mathbb{M}^T + \lambda S_\mu\right| + k\log\left|\tfrac{1}{k}\mathbb{E}\mathbb{E}^T + \lambda S_\epsilon\right| \\ \text{s.t.} \quad & \mathbb{X} = \mathbb{E} + \mathbb{M}\Phi, \end{aligned} \tag{11}$$

where irrelevant constant factors have been removed. Clearly (11) is equivalent to the objective function stated in Theorem 1.

3. Low-Rank Likelihood Ratio Test

Given an unknown image pair x_1 and x_2, it is straightforward to show using block matrix identities that the likelihood ratio statistic required for testing purposes is given by (omitting constant terms)

$$r(x_1, x_2) = x_1^T A x_1 + x_2^T A x_2 + x_1^T B x_2, \tag{12}$$

where

$$\begin{aligned} A &= (T_\mu + T_\epsilon)^{-1} - \left[T_\mu + T_\epsilon - T_\mu (T_\mu + T_\epsilon)^{-1} T_\mu\right]^{-1} \\ B &= \left(T_\mu + \frac{1}{2}T_\epsilon\right)^{-1} T_\mu T_\epsilon^{-1}. \end{aligned} \tag{13}$$

Starting with A, we can use the Woodbury matrix inversion identity to show that

$$\left[T_\mu + T_\epsilon - T_\mu (T_\mu + T_\epsilon)^{-1} T_\mu\right]^{-1} =$$
$$(T_\mu + T_\epsilon)^{-1} - (T_\mu + T_\epsilon)^{-1} T_\mu \left[T_\mu (T_\mu + T_\epsilon)^{-1} T_\mu - T_\mu - T_\epsilon\right]^{-1} T_\mu (T_\mu + T_\epsilon)^{-1}.$$

Therefore, it follows that

$$A = U^{-1} T_\mu \left[T_\mu U^{-1} T_\mu - U\right]^{-1} T_\mu U^{-1}, \tag{14}$$

where $U = T_\mu + T_\epsilon$. From the multiplication by T_μ in (14), it is then obvious that

$$\text{rank}[A] \leq \text{rank}[T_\mu]. \tag{15}$$

Additionally, based on the Shur Complement Lemma,

$$U - T_\mu U^{-1} T_\mu \succeq 0. \tag{16}$$

Therefore, $\left[T_\mu U^{-1} T_\mu - U\right]^{-1}$ is negative semi-definite, symmetric and so A will always be a negative semi-definite, symmetric matrix as well. This implies that $A = -\Phi^T \Phi$ for some Φ such that $\text{rank}[\Phi] \leq \text{rank}[T_\mu]$. Consequently $x^T A x$ can be implemented using a simple low-rank transformation of the image Φx.

Turning now to B, we first define $T_\mu = \Theta\Theta^T$, which is always possible since T_μ is a covariance. Using standard linear algebra identities, we have

$$\begin{aligned} \left(T_\mu + \frac{1}{2}T_\epsilon\right)^{-1} T_\mu T_\epsilon^{-1} &= \left(\Theta\Theta^T + \frac{1}{2}T_\epsilon\right)^{-1} \Theta\Theta^T T_\epsilon^{-1} \\ &= 2T_\epsilon^{-1}\Theta \left(I + \Theta^T \Theta\right)^{-1} \Theta^T T_\epsilon^{-1}. \end{aligned}$$

And so clearly B is a positive semi-definite, symmetric matrix with

$$\text{rank}[B] \leq \text{rank}[\Theta] = \text{rank}[T_\mu]. \tag{17}$$

Analogous to A, it then naturally follows that $x_1^T B x_2$ can be efficiently computed using simple low-rank transformations of the requisite images.

4. More Technical Details of the Proposed EM Algorithm

The proposed EM algorithm from Section 1 computes an approximate E-step, and thus it is more accurately categorized as a generalized EM algorithm. We have claimed that the full E-step could be computed; however, we choose an approximate form for multiple reasons as mentioned above. Here we provide some more details regarding the full algorithm for completeness.

It is easily shown that the full E-step only actually requires that we compute the $d \times d$ block-diagonal elements of $E[HH^T]$ (subject index omitted). Because we have already described the computation of $E[H]$, we need only compute block-diagonal

elements of $Cov[H]$. Similar to the derivation of $E[H]$, and after applying the Woodbury matrix inversion identity, it follows that

$$Cov(H|X) = \Omega - \Omega P^T (P\Omega P^T)^{-1} P\Omega. \tag{18}$$

Moreover, block-diagonal elements of $Cov[H]$ can be computed for all images with even less computational complexity than is required for $E[H]$.

While details will be deferred to a subsequent journal publication, it can be shown (using a related but substantially more complex version of the proof of Theorem 1) that inclusion of this additional covariance factor into the EM algorithm implies that we are now optimizing the alternative cost function

$$\begin{aligned} \min_{\mathbb{M},\mathbb{E};\ \Psi,\Gamma \succeq 0} \quad & \text{trace}\left[\left(\mathbb{M}\mathbb{M}^T + n\lambda S_\mu\right)\Psi^{-1}\right] + \text{trace}\left[\left(\mathbb{E}\mathbb{E}^T + k\lambda S_\epsilon\right)\Gamma^{-1}\right] + \\ & (k-n)\log|\Gamma| + \textstyle\sum_{i=1}^n \log|m_i\Psi + \Gamma| \\ \text{s.t.} \quad & \mathbb{X} = \mathbb{E} + \mathbb{M}\Phi. \end{aligned} \tag{19}$$

The distinction from (6) is found in the log-det factors, which now include a determinant expression that couples the variational parameters Ψ and Γ. Interestingly though, in the limit as Ψ and Γ occupy independent subspaces, these log-det terms collectively converge to $k\,|\Gamma| + n\log|\Psi|$, and so the optimization problems will roughly resemble one another when Γ and Ψ are nearly independent. This is consistent with the fact that if T_μ and T_ϵ occupy occupy different subspaces, then it can be shown (using a series of linear algebra manipulations) that $Cov[H] = 0$, and hence the full and approximate EM updates will be equivalent. However, in the general case, there exists no closed-form solution for Γ and Ψ analogous to (7), and so we cannot simplify the model in the same way as before for analysis purposes.

Regardless, we may ask then what relevant differences exists between (6) and (19), and is there any occasion to prefer one over the other. This is an interesting theoretical question that may ultimately have practical significance, even though in preliminary experimentation we found little distinction. For example, it can be shown that (6) has certain undesirable, degenerate minimizing solutions that occur in the simplified setting where $\lambda \rightarrow 0$. Consequently, given an adversarial initialization, (6) may fail to produce a reasonable solution even when the optimal subspaces are sufficiently low-rank and separable such that closed-form decomposition techniques would suffice. In contrast, this undesirable degeneracy does not exist when using (19) because of a smoothing effect introduced by the coupling between Γ and Ψ in the log-det term. While the details and implications of these distinctions are relevant to a variety of Bayesian face algorithms, we defer further discussion and analysis to a subsequent journal publication.

5. More Samples from the datasets used in our experiments

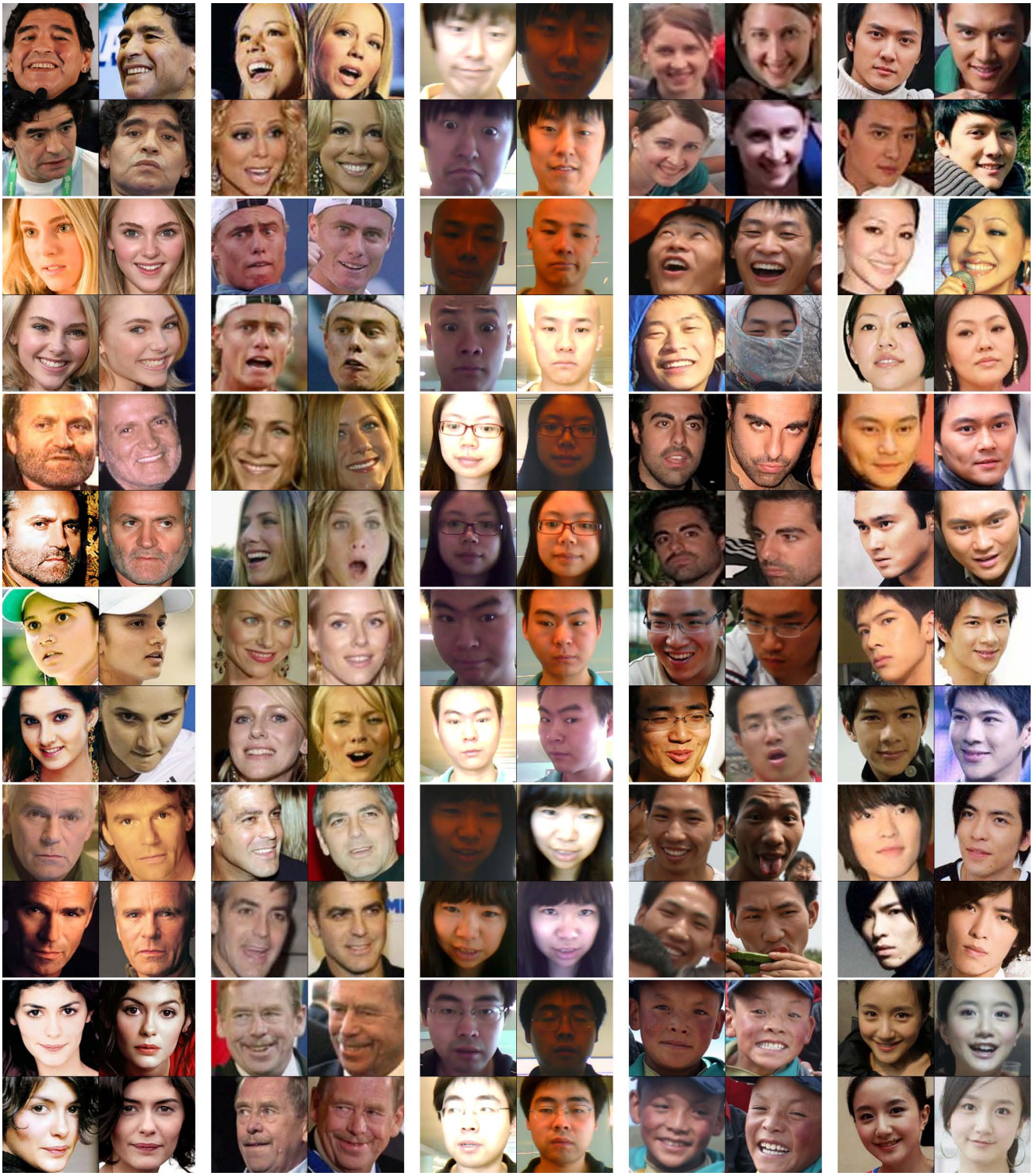

Figure 1. Some more samples of the datasets used in our experiments. From left to right: WDRef, LFW, Video Camera Dataset, Family Album Dataset, and WDAsian.