Synergy-based Learning of Facial Identity

Martin Köstinger, Peter M. Roth, and Horst Bischof

Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{*koestinger,pmroth,bischof*}*@icg.tugraz.at*

Abstract. In this paper we address the problem that most face recognition approaches neglect that faces share strong visual similarities, which can be exploited when learning discriminative models. Hence, we propose to model face recognition as multi-task learning problem. This enables us to exploit both, shared common information and also individual characteristics of faces. In particular, we build on Mahalanobis metric learning, which has recently shown good performance for many computer vision problems. Our main contribution is twofold. First, we extend a recent efficient metric learning algorithm to multi-task learning. The resulting algorithm supports label-incompatible learning which allows us to tap the rather large pool of anonymously labeled face pairs also for face identification. Second, we show how to learn and combine person specific metrics for face identification improving the classification power. We demonstrate the method for different face recognition tasks where we are able to match or slightly outperform state-of-the-art multi-task learning approaches.

1 Introduction

For humans the recognition of familiar faces is straight forward. Computational face recognition matches the performance of humans in controlled environments, however, often fails under unconstrained real-world conditions (e.g., diversity in viewpoint, lighting, clutter, or occlusions). This can be explained by essential differences in human and machine learning. Typically when machine learning techniques learn a specific visual model they focus on individual characteristics and neglect general concepts or visual commonalities of similar objects. In contrast, the human visual system learns in a more synergistic way that benefits from commonalities and takes into account prior knowledge. Hence, for computational recognition systems it would be beneficial also to exploit such information.

One popular concept that addresses this demand is transfer learning, which aims at improving the performance of a target learning task by also exploiting collected knowledge of different sources [1]. Two related aspects are domain adaptation and multi-task learning. Domain adaptation tries to bridge the gap between a source domain with sufficient labeled data to a specific target domain with little or no labels [1]. In contrast, multi-task learning (MTL) [2] approaches a cluster of similar tasks in parallel. Each task describes a target learning problem and contributes labeled data. The knowledge transfer between the tasks is then established through a shared intermediate representation. The basic assumption is that it is easier to learn several hard tasks simultaneously than to learn those isolated. In this way underrepresented tasks that have only a limited number of labeled samples can be handled. Prominent approaches rely on neural nets [3, 4] (sharing layers) or support vector machines [5] (sharing weight vectors).

In this paper, we adapt multi-task learning for real-world, large-scale face recognition. In order to cope with the real-world challenges we want to incorporate as much relevant information as possible. In particular, given by similar/dissimilar labeled face pairs, where we have no access to the actual class labels. These labeled pairs are mainly used for face verification (deciding if two faces match) and are rather easy to obtain also on a large scale. For face identification it is not immediately obvious how to make use of this anonymous information. But these additional face pairs allow us to learn a more robust measure of face similarity. Multi-task learning then spreads this knowledge between the tasks. Hereby, to enable meaningful transfer of knowledge, multi-task learning faces the problem of different label sets. On the one hand side for face identification the label set consists of class labels while on the other hand side we have only equivalence labels. Thus, one important aspect of multi-task learning is label-incompatible learning, the support of different label sets for different learning tasks. Particularly, the successful multi-task adaptation of support vector machines [5] lacks this feature.

Recently, Mahalanobis metric learning [6, 7] showed favorable performance for various computer vision tasks including face verification [8]. The goal is to find a global linear transformation of the feature space such that relevant dimensions for classification or ranking are emphasized while irrelevant ones are discarded. One particular advantage is that Mahalanobis metric learning methods usually operate on the space of pairwise differences, thus enabling label-incompatible learning. The method of Parameswaran and Weinberger [9] extends Mahalanobis metric learning to the multi-task paradigm. Nevertheless, due to the particular optimization it relies on labeled triplets and can thus not benefit from data just labeled with equivalence constraints. Further, it requires computationally expensive iterations making it impractical for large-scale applications. Hence, to capitalize on multi-task learning for face recognition, one faces the additional challenges of scalability and the ability to deal just with equivalence labels.

To meet these requirements, we extend a recent efficient metric learning algorithm [10] to the multi-task paradigm. The resulting algorithm enables label-incompatible learning as it only relies on pairwise equivalence labels. These are considered as natural inputs to distance metric learning algorithms as similarity functions basically establish a relation between pairs of points. In particular, we want to learn specific Mahalanobis distance metrics for each person. This is inspired by the recent finding of Weinberger and Saul [11] that especially for large-scale applications better results can be obtained by learning multiple distance metrics. Also many other learning algorithms cast a complex multi-class problem in series of simpler, often two class, problems, followed by a voting rule to form the final decision [12]. Thus, inspired by the successful strategy applied for multi-class support vector machines we intend to learn individual distance metrics. Our method is scalable to large datasets and not prone to over-fitting. To demonstrate the merits of our method we compare it to recent multi-task and metric learning approaches on the challenging PubFig [13] face recognition benchmark.

2 Multi-Task Metric Learning for Face Recognition

In the following, we introduce our new multi-task metric learning approach for face recognition. First, in Section 2.1 we briefly describe the metric learning approach introduced in [10], which is very efficient in training as it avoids complex iterative computations and is thus scalable to large datasets. Next, in Section 2.2, we extend this approach for the multi-task domain. Finally, in Section 2.3 we introduce a voting scheme that allows for classification using multiple metrics. The overall goal is to combine several person specific metrics to a multi-class decision which should lead to lower error rates.

2.1 Mahalanobis Metric Learning

One prominent approach for metric learning is to learn a Mahalanobis distance $d^2_{\mathbf{M}}$, which measures the squared distance between two data points $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$:

$$d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j) \; . \tag{1}$$

The only requirement to induce a valid (pseudo) metric is that $\mathbf{M}$ is a symmetric positive semi-definite matrix. Several different approaches (e.g., [11], [7], or [8]) have been proposed that address different loss functions or regularizations to optimize such a metric for specific problems. However, such approaches typically require complex iterative, computationally expensive optimization schemes and fully labeled data. Instead, *KISS metric* learning (KISSME) [10] overcomes these limitations by introducing an efficient statistical motivated formulation that allows to learn just from equivalence constraints. Analog to the KISS principle (*keep it simple and straightforward!*) the method is conceptually simple and efficient per design.

For the following discussion let $\mathbf{x}_i$,$\mathbf{x}_j \in \mathbb{R}^d$ be a pair of samples and $y_i, y_j \in \{1, 2, \dots, c\}$ the according labels. Further we define a set of similar pairs $\mathcal{S} = \{(i,j)\,|y_i = y_j\}$ and a set of dissimilar pairs $\mathcal{D} = \{(i,j)\,|y_i \neq y_j\}$. The goal of KISSME is to decide whether a pair (i,j) is similar or not. From a statistical inference point of view the optimal statistical decision can be obtained by a likelihood ratio test. Hereby, the hypothesis H_0 that the pair is dissimilar is tested against H_1 that the pair is similar:

$$\delta(\mathbf{x}_{ij}) = \log\left(\frac{p(\mathbf{x}_{ij}|H_0)}{p(\mathbf{x}_{ij}|H_1)}\right) = \log\left(\frac{f(\mathbf{x}_{ij}|\theta_0)}{f(\mathbf{x}_{ij}|\theta_1)}\right) \; , \tag{2}$$

where δ is the log-likelihood ratio, $f(\mathbf{x}_{ij}|\theta)$ is a pdf with parameters θ and $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$. Thus, KISSME casts the metric learning problem into the space of pairwise differences, as also the similarity Eq. (1) is defined via pairwise differences. This space has zero-mean and is invariant to the actual locality of the samples in the feature space. Assuming zero-mean Gaussian distributions within the difference space Eq. (2) can be re-written to

$$\delta(\mathbf{x}_{ij}) = \log \left(\frac{\frac{1}{\sqrt{2\pi|\Sigma_{\mathcal{D}}|}} \exp(-1/2\ \mathbf{x}_{ij}^T\ \Sigma_{\mathcal{D}}^{-1}\ \mathbf{x}_{ij})}{\frac{1}{\sqrt{2\pi|\Sigma_{\mathcal{S}}|}} \exp(-1/2\ \mathbf{x}_{ij}^T\ \Sigma_{\mathcal{S}}^{-1}\ \mathbf{x}_{ij})} \right), \tag{3}$$

where $\Sigma_{\mathcal{S}}$ and $\Sigma_{\mathcal{D}}$ are the covariance matrices of $\mathcal{S}$ and $\mathcal{D}$, respectively. Let $\mathbf{C}_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$ be the outer product of the pairwise differences of $\mathbf{x}_i$ and $\mathbf{x}_j$, the covariance matrices can be written as

$$\Sigma_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{(i,j)\,\in\,\mathcal{S}} \mathbf{C}_{ij}\ , \quad \Sigma_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{(i,j)\,\in\,\mathcal{D}} \mathbf{C}_{ij}\ . \tag{4}$$

The maximum likelihood estimate of the Gaussian is equivalent to minimize the distances from the mean in a least squares manner. This allows KISSME to find respective relevant directions for $\mathcal{S}$ and $\mathcal{D}$. By taking the log and discarding the constant terms we can simplify Eq. (3) to

$$\delta(\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\ \Sigma_{\mathcal{S}}^{-1}\ \mathbf{x}_{ij} - \mathbf{x}_{ij}^T\ \Sigma_{\mathcal{D}}^{-1}\ \mathbf{x}_{ij} = \mathbf{x}_{ij}^T(\Sigma_{\mathcal{S}}^{-1} - \Sigma_{\mathcal{D}}^{-1})\mathbf{x}_{ij}\ . \tag{5}$$

Finally, the Mahalanobis distance matrix $\mathbf{M}$ is obtained by

$$\mathbf{M} = \left(\Sigma_{\mathcal{S}}^{-1} - \Sigma_{\mathcal{D}}^{-1}\right)\ . \tag{6}$$

2.2 Multi-Task Metric Learning

Now having introduced KISSME, we can extend formulation Eq. (6) to the multi-task learning paradigm. The general idea of multi-task learning is to consider T different, but related learning tasks in parallel. In our case a task is to learn a face verification model for a specific person, and the relation is intuitively given via the shared visual properties of faces. There are different concepts to realize such a setting. In particular, we adopt the formulation of Parameswaran and Weinberger [9]. We model the individual metric for each task $t \in \{1, 2, \ldots, T\}$ as combination of a shared metric $\mathbf{M}_0$ and a task-specific metric $\mathbf{M}_t$:

$$d_t^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{M}_0 + \mathbf{M}_t)(\mathbf{x}_i - \mathbf{x}_j)\ . \tag{7}$$

Each task defines a subset of task specific samples given by the index set $\mathcal{I}_t$. Hence, to adopt the formulation Eq. (7) for the KISS metric, we have to define a task-specific subset of similar and dissimilar sample pairs: $\mathcal{S}_t = \{(i,j) \in \mathcal{I}_t | y_i = y_j\}$ and $\mathcal{D}_t = \{(i,j) \in \mathcal{I}_t | y_i \neq y_j\}$. In cases where $\mathcal{S}_t$ and $\mathcal{D}_t$ is not given, these sets can be sampled randomly of the actual class labels. Hence, according to Eq. (6) we can estimate task specific metrics by

$$\mathbf{M}_t = \left(\frac{1}{|\mathcal{S}_t|} \sum_{(i,j) \,\in\, \mathcal{S}_t} \mathbf{C}_{ij} \right)^{-1} - \left(\frac{1}{|\mathcal{D}_t|} \sum_{(i,j) \,\in\, \mathcal{D}_t} \mathbf{C}_{ij} \right)^{-1} . \tag{8}$$

Similarly, by estimating the weighted sum over the individual task specific characteristic we get the shared or common metric

$$\mathbf{M}_0 = \left(\frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\mathcal{S}_t|} \sum_{(i,j) \,\in\, \mathcal{S}_t} \mathbf{C}_{ij} \right)^{-1} - \left(\frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\mathcal{D}_t|} \sum_{(i,j) \,\in\, \mathcal{D}_t} \mathbf{C}_{ij} \right)^{-1} . \tag{9}$$

Then, the final individual Mahalanobis distance metric is given by

$$\hat{\mathbf{M}}_t = \mathbf{M}_0 + \mu \; \mathbf{M}_t \; . \tag{10}$$

Intuitively, $\mathbf{M}_0$ picks up general trends across all tasks and thus models commonalities. In contrast, $\mathbf{M}_t$ models task-specific characteristics. As only free parameter we retain a balancing factor μ between the task specific metric $\mathbf{M}_t$ and the shared metric $\mathbf{M}_0$. Intuitively, the more samples a task contributes the more focus lies on its specific metric.

2.3 Multi-Task Voting

To fully exploit the power of our multi-task metric learning method for face recognition, we combine multiple, person specific, metrics into a multi-class decision. However, the outputs of the different metrics are not necessarily compatible and cannot be compared directly. A prominent strategy to reconcile classifier outputs is to calibrate them by fitting a sigmoid curve to a held-out set [14]. Nevertheless, since such an approach requires a large amount of labeled data, it is inapplicable for our purpose. Another successful strategy is to assign the class that wins most pairwise comparisons [15], also referred as *max-wins* rule.

To adapt this strategy for multi-task metric learning, we assume that the positive samples for task t coincidence with the class label $\mathbf{x}_i : y_i = t$. Then the combination rule

$$\begin{aligned} &\arg\max_t(\mathbf{x}_i) = \\ &\arg\max_t \sum_{u \neq t} \Bigg[\mathrm{I} \left(\min_{j \in \mathcal{I}_t \wedge y_j = t} d_t^2(\mathbf{x}_i, \mathbf{x}_j) \leq \min_{k \in \mathcal{I}_u \wedge y_k = u} d_t^2(\mathbf{x}_i, \mathbf{x}_k) \right) \\ &\qquad + \mathrm{I} \left(\min_{j \in \mathcal{I}_t \wedge y_j = t} d_u^2(\mathbf{x}_i, \mathbf{x}_j) \leq \min_{k \in \mathcal{I}_u \wedge y_k = u} d_u^2(\mathbf{x}_i, \mathbf{x}_k) \right) \Bigg] \end{aligned} \tag{11}$$

checks if the minimum distance of a given test sample $\mathbf{x}_i$ to class t is smaller than to class u. The indicator function

$$\mathrm{I}(x) = \begin{cases} 1 \text{ if } x \text{ is true} \\ 0 \text{ otherwise} \end{cases} \tag{12}$$

scores for class t if this is true. This comparison is done with the individual distance metric of task t. Further, we also compare the distances under the complementary distance metric of task u. The basic idea is that if class t scores even under that metric it is an indicator for class t. Intuitively, the final decision is for the class that wins most pairwise comparisons.

3 Experiments and Evaluations

In the following, we demonstrate the performance of our method on the Public Figures Face Database (PubFig) [13]. The dataset can be considered as very challenging as it exhibits huge variations in pose, lighting, facial expression and general imaging and environmental conditions. As features we use the "high-level" description of visual face traits [13], which describes the presence or absence of 73 visual attributes, such as gender, race, hair color etc. For the intended face identification benchmark we organize the data similar to the existing verification protocol in 10 folds for cross-validation. Therefore, we split the images of each individual into 10 disjoint sets. The goals of our experiments are twofold. First, in Section 3.1 we show that multi-task learning allows us to successfully exploit additional data with anonymous pairwise labels for face identification. Next, in Section 3.2 we show that multi-task learning of person specific metrics boosts the performance for face identification. In particular, we show that the power lies in the combination of multi-task learning and the person specific metrics, as it is not sufficient to learn them off-the-shelf. Further, we compare our results to standard metric learning and related multi-task learning approaches.

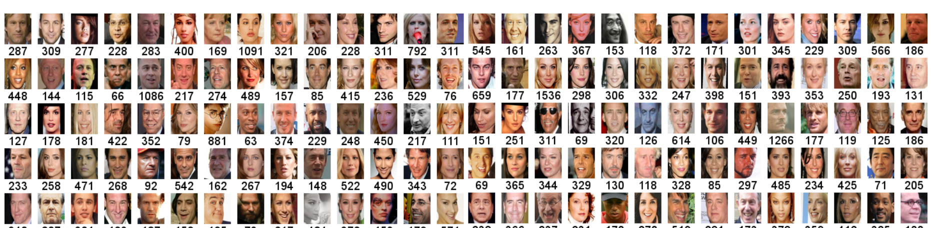
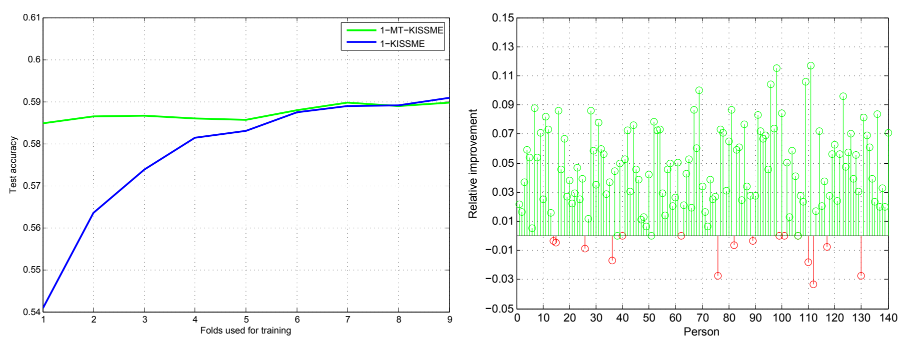


Fig. 1: PubFig database [13]: The evaluation set contains 42,461 images of 140 individuals. The number of images per individuals ranges from 63 (Dave Chappelle) to 1536 (Lindsay Lohan). We split the images in 10 non-overlapping folds for cross-validation.

3.1 Inducing Knowledge from Anonymous Face Pairs to Face Identification

First, we show that multi-task learning allows us to transfer general knowledge about face similarity from anonymous face pairs to face identification. In order to enable a meaningful transfer of knowledge hereby multi-task learning faces the problem of different label sets. We test a multi-task learning scenario with two learning tasks, one with pairwise equivalence labels for the face pairs and one with class labels for face identification. The goal is to show that the additional anonymous face pairs help to improve the face identification performance. We sample the pairs randomly of the predefined development split of the dataset, containing 60 people. For the identification task we use the evaluation set, containing 140 people (Fig. 1). Thus, we ensure that the subjects for the tasks are mutually exclusive. For a given test sample we perform k-NN classification using a single metric to the 140 classes. Using different values for k revealed that there is no significant performance change, although simple nearest neighbor assignment leads to the best performance. Thus, we stick to a simple nearest neighbor assignment.



(a) (b)

Fig. 2: Benefiting from additional pairwise labels for face identification on the PubFig dataset: (a) k-NN classification accuracy of KISSME multi-task vs. standard single-task learning in relation to the amount of training data; (b) relative performance change per person from single-task to multi-task learning after using one fold for training. Green indicates positive induction while red indicates a negative induction.

In Figure 2 (a) we plot the face identification performance in relation to amount of data used to train the metric. Testing is done on a held-out set via 10 fold cross-validation. In each step we increase the number of folds used to train the identification task by one. As expected, the distance metric trained via multi-task learning (1-MT-KISSME) yields reasonable results right from the

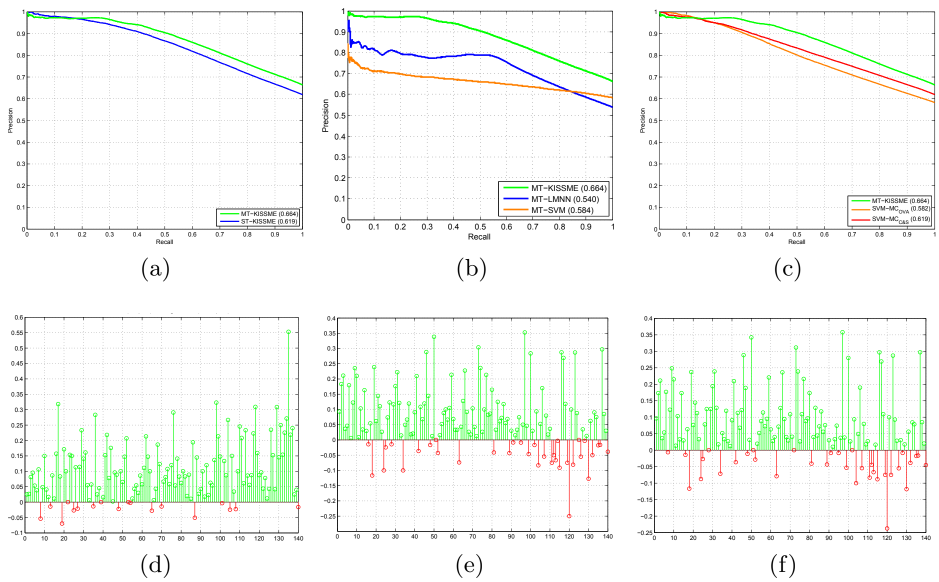


Fig. 3: PubFig face identification benchmark. Comparison of the proposed method (MT-KISSME) to (a) single-task learning , (b) to other MTL methods, and (c) to SVMs. Numbers in parentheses denote the precision of the respective method at full recall. Bottom row, (d)-(f), compares the accuracy per person of the best performing competing method of the plot above to MT-KISSME.

beginning. Obviously, it is able to reuse knowledge of the anonymous face pairs. In contrast, the distance metric trained without the additional pairwise labels (1-KISSME) needs by far more data to reach the same performance. In Figure 2 (b), we compare the relative performance change per person from standard single-task learning to multi-task learning, after one training fold. In most cases an improvement can be obtained.

3.2 Person specific Metric Learning

Second, we demonstrate the performance of our MTL method to learn person specific distance metrics. To show the merit of our method we compare it to recent MTL methods [5, 9] and also benchmark to multi-class support vector machines [16, 17]. We report the face identification performance in a refusal to predict style. Therefore, we rank and threshold the classifier scores. In that sense, recall means the percentage of samples which have a higher score than the current threshold and thus are labeled. Precision means the ratio of correctly labeled samples.

In Figure 3 (a) we compare, as a sanity check, the performance of estimating person specific metrics via multi-task vs. single-task learning. The MTL method outperforms the single-task learning over most levels of recall. At full recall the performance difference is about 4.5%. The main advantage of our MTL method is revealed if we compare the recognition accuracy per person. With multi-task learning we reach a person accuracy of 63.10% while single-task reaches only 54.08%. Thus, it is favorable to learn person specific metrics multi-task. In Figure 3 (d) we compare the relative performance change per person. Only for a small number of classes the performance drops slightly while for the vast number the performance increases.

Next, in Figure 3 (b) we benchmark to recent MTL methods, MT-LMNN [9] and MT-SVM [5]. Both methods are not really able to capitalize on the synergies of the face identification task. Both methods are outperformed by MT-KISSME over all levels of recall. At full recall the respective performance gain compared to MT-LMNN is 12.4%, compared to MT-SVM 8%. In Figure 3 (e) we plot the relative performance change on person level compared to MT-SVM. Hence, our method is able also to compete with two recent MTL approaches. Compared to the MT-SVM one advantage may be that MT-KISSME operates in the space of pairwise differences, which eases meaningful transfer of knowledge between the learning tasks. Further, compared to both competing MTL methods MT-KISSME is able to gain information from pairwise labels.

Finally, in Figure 3 (c) we benchmark our method to multi-class support vector machines. Particularly, the method of Crammer and Singer [16] has shown recent success also compared to metric learning methods [6]. The standard multi-class one-vs-all SVM reaches with 58.4% at full recall about the same performance as the MT-SVM. The method of Crammer and Singer [16] beats this by 3.7%. This may be accounted to the fact that it attempts to solve a single multi-class optimization problem that is better suited for unbalanced datasets. Nevertheless, MT-KISSME outperforms the one-vs-all method by 8.5% and the method of Crammer and Singer by 4.5%.

4 Conclusion

In this work we presented a synergistic approach to exploit shared common as well as person specific information for face recognition. By extending KISSME [10] metric learning we developed a multi-task learning method that is able to learn from just equivalence constraints, thus, enabling label-incompatible learning. Overall, we get a conceptually simple but very effective model, which is scalable to large datasets. Further, we showed that learning person specific metrics boosts the performance for face identification. In particular, we revealed that the power lies in the combination of multi-task learning and person specific metrics, as it is not sufficient to learn the metrics decoupled. To show the merits of our method we conducted two experiments on the challenging large-scale PubFig face benchmark. We are able to match or slightly outperform recent multi-task learning methods and also multi-class support vector machines.

Acknowledgments. The work was supported by the Austrian Science Foundation (FWF) project Advanced Learning for Tracking and Detection in Medical Workflow Analysis (I535-N23) and by the Austrian Research Promotion Agency (FFG) project SHARE in the IV2Splus program.

References

1. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering **22** (2010) 1345–1359
2. Caruana, R.: Multitask learning: A knowledge-based source of inductive bias. In: Proc. IEEE Intern. Conf. on Machine Learning. (1993)
3. Caruana, R.: Multitask learning. Machine Learning **28** (1997) 41–75
4. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proc. IEEE Intern. Conf. on Machine Learning. (2008)
5. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proc. Intern. Conf. on Knowledge discovery and data mining. (2004)
6. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems. (2006)
7. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proc. IEEE Intern. Conf. on Machine Learning. (2007)
8. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? Metric learning approaches for face identification. In: Proc. IEEE Intern. Conf. on Computer Vision. (2009)
9. Parameswaran, S., Weinberger, K.: Large margin multi-task metric learning. In: Advances in Neural Information Processing Systems. (2010)
10. Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: Proc. IEEE Intern. Conf. on Computer Vision and Pattern Recognition. (2012)
11. Weinberger, K.Q., Saul, L.K.: Fast solvers and efficient implementations for distance metric learning. In: Proc. IEEE Intern. Conf. on Machine Learning. (2008)
12. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. Journal of Machine Learning Research **5** (2004) 101–141
13. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and Simile Classifiers for Face Verification. In: Proc. IEEE Intern. Conf. on Computer Vision. (2009)
14. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large-Margin Classifiers. MIT Press (1999)
15. Friedman, J.H.: Another approach to polychotomous classification. Technical report, Department of Statistics, Stanford University (1996)
16. Crammer, K., Singer, Y., Cristianini, N., Shawe-taylor, J., Williamson, B.: On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research **2** (2001) 265–292
17. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. on Intelligent Systems and Technology **2** (2011) 27:1–27:27