Efficient Face Search using Coordinated Local Metric Learning

Shreyas Saxena Jakob Verbeek
LEAR Team, INRIA Grenoble - Rhône-Alpes, France

1. Introduction

Metric learning is used to obtain distance measures that are meaningful to a certain task at hand. It has many applications in computer vision, *e.g.* in local descriptor matching [5], fine-grained object comparison [15], and face verification [7]. The latter is important for forensics to match face images acquired from, *e.g.*, surveillance cameras to a database of faces with known identity. Forensic face recognition is extremely challenging since images are acquired in a setting where variations in pose and expression can not be controlled. Metric learning suppresses effects due to such nuisance factors in face signatures, so that better matching results can be obtained.

Most work on metric learning considers supervised learning of Mahalanobis metrics, see *e.g.* [4, 6, 7, 13, 19]. Supervision is typically given by positive and negative pairs that should be close and far apart respectively according to the learned metric. Mahalanobis metrics are equivalent to the ℓ_2 metric after linear projection of the data. For complex data distributions, however, linear projection of the data might not be sufficient to obtain a suitable metric. To overcome this restriction non-linear projections can be obtained using the "kernel trick" [6], by learning (convolutional) neural networks [3], or by learning local Mahalanobis metrics that each operate in a different part of the input space [1, 2, 8, 10, 14, 16, 18, 19, 22].

Most existing local metric learning approaches, however, do not allow to compute distances between points assigned to different clusters, or distances are defined in an asymmetric manner. Unlike global metric learning, they can not be interpreted as computing the ℓ_2 distance after a linear projection of the data, which hinders data visualization, and the use of efficient ℓ_2-based retrieval techniques such as product quantization and multiple-assignment retrieval [11].

The contribution of our approach is that we embed local metrics in a global representation, in which the ℓ_2 metric can be used. This allows us (i) to compute distances between points regardless to which local cluster they belong, (ii) visualize data in a single view, and (iii) use efficient ℓ_2-based retrieval methods.

In this abstract we present our approach and a selection of experimental results on the "Labeled Faces in the Wild" dataset [9], the *de facto* standard dataset for uncontrolled face recognition. Results show that our approach improves over previous local and global metric learning methods.

2. Coordinated Local Metric Learning

Since any positive definite $D \times D$ matrix M can be decomposed as $M = L^\top L$, the Mahalanobis distance between two points x_i and x_j can be written as the ℓ_2 distance between these points after projection with L, *i.e.*

$$(x_i - x_j)^\top M (x_i - x_j) = \| Lx_i - Lx_j \|_2^2 . \tag{1}$$

To obtain a more general class of metrics, we define multiple local Mahalanobis metrics. We cluster the data using a k-component Gaussian mixture model (GMM), and learn k separate metrics associated with the clusters. We can compute distances between points assigned to the same cluster s using a local metric learned for that cluster, defined by a local projection matrix L_s.

The local metrics can be interpreted as the ℓ_2 distances after locally mapping the data points x to different, local, coordinate systems via projections $L_s x$. Since the ℓ_2 metric is invariant to translation, rotation, and reflection of the coordinates, we can arbitrarily modify the projection of x_i to a local coordinate system to

$$z_{is} := R_s L_s x_i + b_s, \tag{2}$$

where R_s denotes an orthonormal matrix, *i.e.* for which $R_s^\top R_s = I$, which can implement rotations and reflections, and b_s denotes a translation. Using these transformation we can coordinate the local projections so that they align across different local models.

To learn both local metrics and their alignment parameters in a joint manner, we absorb R_s into L_s without loss of generality. We map the data points x_i to a global coordinate system using the weighted average of the projections obtained using the different local models

$$z_i := \sum_{s=1}^{k} q_{is} z_{is} = \sum_{s=1}^{k} q_{is} \big(L_s x_i + b_s \big), \tag{3}$$

# local metrics	LBP	FV	CNN
$k = 4$	44.02	75.18	61.43
$k = 16$	49.00	76.20	66.07
$k = 32$	49.98	75.61	70.98
Cross validated k	49.89 (26)	74.99 (28)	70.98 (32)
Global LDML metric	36.95	68.12	58.46
ℓ_2 metric	13.24	22.88	63.06

Table 1. Retrieval mAP while varying the number of local metrics.

where q_{is} is the GMM soft-assignment of x_i to cluster s. In the global coordinates given by the z_i we can compare any pair of points, regardless of whether they are assigned to different clusters or not.

Our goal is now to learn $\{L_s, b_s\}$ so that z_i and z_j are close for positive pairs, and far away for negative pairs for a given fixed clustering. Note that we can rewrite this weighted average of locally linear projections as a single linear projection

$$z_i = \tilde{L}\tilde{x}_i, \tag{4}$$

where $\tilde{L} := (L_1, b_1, \ldots, L_k, b_k)$ collects the local linear projections, and $\tilde{x}_i := \left(q_{i1}(x_i^\top, 1), \ldots, q_{ik}(x_i^\top, 1)\right)^\top$ contains k copies of x_i appended with a one, each weighted by the corresponding soft-assignment. Since the z_i are a linear projection of the transformed input vectors $\tilde{x}_i$, the ℓ_2 distance between the z_i's is equivalent to a Mahalanobis distance between the $\tilde{x}_i$'s. Therefore, learning a globally aligned ensemble of local metrics is similar to learning a global Mahalanobis metric. The difference is that we use the expanded data representation given by the $\tilde{x}_i$, which is k times higher dimensional than the original data representation. In practice the $\tilde{x}_i$ do not need to be explicitly stored in memory, by using Eq. (3) to compute the z_i directly from the x_i. The computational cost also grows sub-linearly in k, since the soft-assignments are typically sparse.

In practice, we use the LDML [7] objective function to learn our coordinated local metrics parameterized by $\tilde{L}$. Let the label $y_{ij} \in \{-1, +1\}$ denote whether (x_i, x_j) is a positive or a negative pair. LDML then minimizes the log-loss

$$\mathcal{L}(\tilde{L}, b) = \sum_{i,j} \ln\left\{1 + \exp\left(-y_{ij}(b - ||z_i - z_j||^2)\right)\right\}, \tag{5}$$

where b is a scalar (estimated along with $\tilde{L}$) that determines at which distance pairs are considered positive or negative.

3. Experimental Evaluation

We report results for the Labeled Faces in the Wild (LFW) [9] dataset, which is the most widely used one for uncontrolled face recognition. It contains 13,233 faces of 5,749 people collected from the web. We use the retrieval-based evaluation protocol of Bhatterai *et al*. [1]. There are 423 queries, one for each person in LFW with five or more images. We report the mean average precision (mAP) and the 1-call@n measure which gives the fraction of queries for which at least one of the top n results is correct.

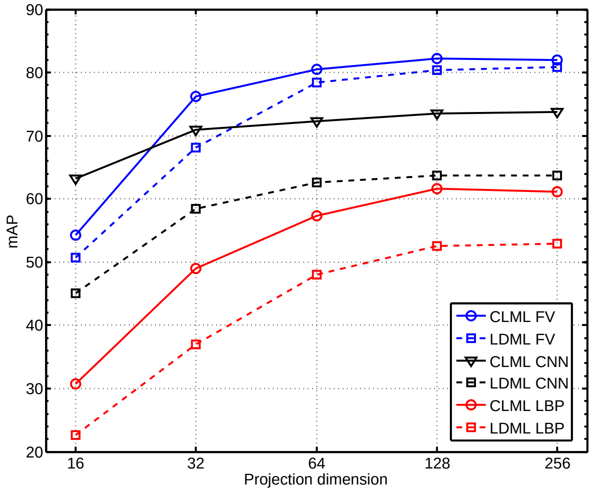

Figure 1. Performance in mAP of local and global metrics for FV, CNN, and LBP features, using different projection dimensions.

We consider three face representations: the LBP feature of Bhattarai *et al*. [1] (9,860 dimensional), the Fisher vector (FV) feature of Simonyan *et al*. [17] (16,896 dimensional), and the penultimate layer of a convolutional neural network that was learned on the CASIA WebFace dataset [21] (320 dimensional).

3.1. Experimental evaluation results

For comparability with [1], we use matrices L_s that project to $d = 32$ dimensions unless stated otherwise.

In Table 1 we evaluate our coordinated local metric learning (CLML) approach while varying the number of local metrics. We also state results when cross-validating the number of local metrics, and compare to a global LDML Mahalanobis metric and the ℓ_2 metric. The results lead to the following observations. (i) CLML generally improves when using more local metrics. (ii) Cross-validation over the number of local metrics successfully selects a (near) optimal number of local metrics. (iii) For all tested settings, CLML consistently improves over global Mahalanobis metrics learned with LDML. Below we cross-validate the number of local metrics for CLML.

In Figure 1 we compare CLML and LDML across a range of projection dimensions. The results show that CLML consistently improves over LDML for all projection dimensions and features. The improvements are particularly large for the CNN and LBP features.

In Figure 2 we compare to the local metric learning results of Bhattarai *et al*. [1], using the LBP features and

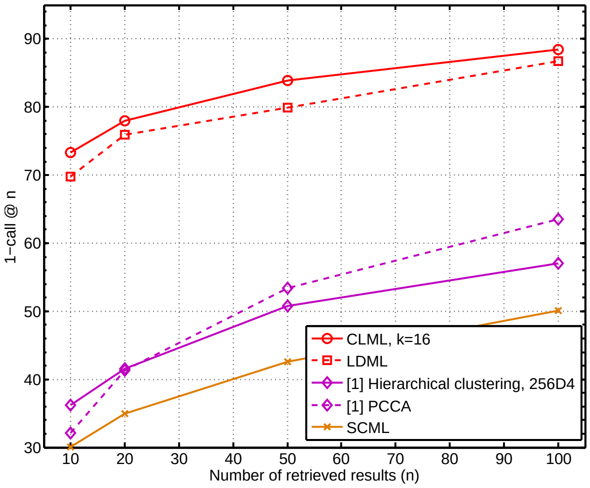

Figure 2. Retrieval on LFW dataset, the results marked with [1] correspond to those reported therein. Results for SCML and LDML have been produced using publicly available code.

$d = 32$ as before for direct comparability. We report the results for their "256D4" setting, which they found to give best results and uses eight local metrics, and also include their global metric learning results obtained with PCCA [13]. We also compare to the recent state-of-the-art SCML approach [16].[1]

Our CLML results substantially improve over the results of Bhatterai *et al*., *e.g*. from under 40% to over 70% for $n = 10$ (Figure 2). The improvement of the global LDML metric over the global PCCA metric of is in part due to the ℓ_2 regularization that was not used by Bhatterai *et al*.

SCML [16] obtains the worst retrieval results. This is because SCML learns metrics that are a linear combination of a limited set of rank-1 base metrics, which is detrimental for high-dimensional data.

Additional experiments, not presented here, show that efficient ℓ_2-based multiple-assignment retrieval [11] can speedup our approach by a factor 16 without loss in performance, and upto a factor 100 for a loss of 4 mAP points. Bhatterai *et al*. report a factor 10 speedup compared to exhaustive search using their hierarchical clustering approach, but obtain substantially worse results. These results, as well as results on the YouTube Faces video dataset [20] and the MOBIO mobile-phone video dataset [12] will be presented at the workshop.

References

[1] B. Bhattarai, G. Sharma, F. Jurie, and P. Pérez. Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval. In *ECCV Workshops*, 2014.

[2] J. Bohné, Y. Ying, S. Gentric, and M. Pontil. Large margin local metric learning. In *ECCV*, 2014.

[3] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

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

[5] A. Dosovitskiy, J. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NIPS*, 2014.

[6] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *NIPS*, 2006.

[7] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. In *ICCV*, 2009.

[8] Y. Hong, Q. Li, J. Jiang, and Z. Tu. Learning a mixture of sparse distance metrics for classification and dimensionality reduction. In *ICCV*, 2011.

[9] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[10] Y. Huang, C. Li, M. Georgiopoulos, and G. Anagnostopoulos. Reduced-rank local distance metric learning. In *ECML*, 2013.

[11] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *PAMI*, 33(1):117–128, 2011.

[12] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matějka, J. Černocký, N. Poh, J. Kittler, A. Larcher, C. Lévy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes. Bi-modal person recognition on a mobile phone: using mobile phone data. In *IEEE ICME Workshop on Hot Topics in Mobile Mutlimedia*, 2012.

[13] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012.

[14] Y.-K. Noh, B.-T. Zhang, and D. Lee. Generative local metric learning for nearest neighbor classification. In *NIPS*, 2010.

[15] E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. In *CVPR*, 2007.

[16] Y. Shi, A. Bellet, and F. Sha. Sparse compositional metric learning. In *AAAI*, 2014.

[17] K. Simonyan, O. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, 2013.

[18] J. Wang, A. Kalousis, and A. Woznica. Parametric local metric learning for nearest neighbor classification. In *NIPS*, 2012.

[19] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009.

[20] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011.

[21] D. Yi, Z. Lei, S. Liao, and S. Li. Learning face representation from scratch. In *Arxiv preprint*, 2014.

[22] D.-C. Zhan, M. Li, Y.-F. Li, and Z.-H. Zhou. Learning instance specific distances using metric propagation. In *ICML*, 2009.

[1] Using code from http://mloss.org/software/view/553.