The Translation Invariant Solution to Quadratic Metric Learning

SUN Bing[1, a], FENG Jufu[1,b] and SHEN Beilun[2,c]

[1]Key Laboratory of Machine Perception (Ministry of Education), School of Electronics Engineering and Computer Science, Peking University, Yiheyuan Road 5, 100871, Beijing, P.R.China.

[2]Chinaoly Co. Ltd., Hangzhou, Zhejiang, P.R.China

[a]bsun@pku.edu.cn, [b]fjf@cis.pku.edu.cn, [c]shenbl@chinaoly.com

Keywords: metric learning, PCA, LDA, translation invariant, image metric

Abstract. Metric learning has drawn great interests for the last decade in the field of computer vision and machine learning. In this paper, we address the importance of the translation invariance (TI) of a metric. Intuitively, translation invariance should be a fundamental requirement for any reasonable image metric, but few metric learning or subspace methods are aware of the TI property when dealing with images. We propose to solve the quadratic metric learning problem in the transform domain based on the result of [1]. The solution is guaranteed to be translation invariant. The TI assumption simplifies the optimization problems considerably. Specifically, it reduces the number of optimization parameters from of O(n2) to O(n); it turns the semi-definite constraint on a matrix to a bound constraint on a function; it applies to multi-dimensional data without having to be stacked to vectors. Experimental results show that the TI solutions are generally on par with the non-TI solutions. Specifically, the TI solutions usually favor small sample size and small reduction dimension. The framework proposed for quadratic metric learning and linear subspace method seems quite promising.

Introduction

Metric learning has drawn great interests for the last decade in the field of computer vision and machine learning since a proper choice of a metric is crucial wherever comparing two objects numerically. Despite that Fukunaga and Flick published their pioneer work in 1984 [2], the term distance metric learning is due to Xing et al. [3], and learning a distance metric from data has been actively studied since then [1,3,4,5].

Most existing metric learning methods assume that the distance metric is in the form of

$$d_G(x, y) = (x - y)^T G(x - y), \tag{1}$$

where x, y are vectors of p×1 and G is a positive semi-definite matrix of p×p. $d_G(\cdot, \cdot)$ is often referred to as the Mahalanobis distance in the literature of metric learning, but the exact meaning of "Mahalanobis distance" is defined by $G=\Sigma^{-1}$, where Σ is the covariance matrix of data. To avoid the ambiguity, $d_G(\cdot, \cdot)$ is referred to as a quadratic distance metric throughout this paper, following the convention in [2].

It is well known that the quadratic metric learning and the linear subspace method are connected: learning a linear transform matrix H is equivalent to learning a quadratic metric matrix G. That is, for any positive semi-definite matrix G > 0, there exists a not unique transform matrix H such that $G = H^TH$ and hence

$$(x - y)^TG(x - y) = (x - y)^TH^TH(x - y) = (H(x - y))^T(H(x - y)). \tag{2}$$

Note that H is not necessarily a square matrix.

On the other hand, Wang et al. proposed an interesting distance metric for images, called the IMage Euclidean Distances (IMED) [6], which is designed a priori rather than learning from data. IMED is invariant to image translation, namely, if the same image translation is applied to two images, their IMED remains invariant. Intuitively, translation invariance (TI) should be a fundamental requirement for any reasonable image metric, but few metric learning or linear subspace methods are aware of the TI property when dealing with images.

Sun et al. [1] constructed a TI transform H such that $G=H^TH$ for IMED and further showed that any TI metric can be applied in the transform domain as a dot product. This is a straight corollary of the Bochner's Theorem [7], however, it does provide some interesting insights. In literature, the metric learning problem is usually formulated as an optimization problem. In order to learn a metric matrix G, there are $p_1^2 \times p_2^2$ elements to be optimized for images of size $p_1 \times p_2$. Moreover, the positive semi-definite constraint on G makes it not always easy to find an efficient algorithm to solve the problem [3,4].

The translation invariant assumption on G is not only intuitively necessary, but also greatly simplifies the problem of metric learning. The metric learning problem under the TI assumption has kp_1p_2 parameters to optimize, which are the samples of $g(\omega)$. (k is the sampling density.) The semi-definite constraint of $G>0$ is reduced to a bound constraint $g(\omega)>0$, thus no semi-definite programming will be encountered as in [3,4].

In this paper, we propose to solve the problem of quadratic metric learning and linear feature extraction in the transform domain. The solution is guaranteed to be translation invariant. The rest of this paper is organized as follows: section 2 presents the translation invariant solutions to some selected quadratic metric learning and linear subspace methods. Section 3 presents experimental results. Finally, a conclusion is given in Section 4.

The translation invariant solutions

The authors in [1] proved that if the metric matrix G is translation invariant, i.e., there exists a sequence g[i] such that $G = g[i - j]$ (G with such a property is referred to as a Toeplitz matrix [8]), then the matrix-vector production can be reduced to simple inner product. As mentioned in section 1, it introduces great simplifications to the optimization problem of metric learning, e.g., O(n) versus $O(n^2)$ parameters, $g(\omega)>0$ bound constraint versus $G>0$ matrix semi-definite constraint. Despite those, another useful property of (2) is that it can be applied to multi-dimensional data without modification.

Formally, if x is multi-dimensional data of $p_1 \times p_2 \times \cdots \times p_d$ and G is a metric tensor with the appropriate size [9], the above statement remains true with the multi-dimensional DTFT [1]. This property is quite neat since algorithms can be applied to multi-dimensional case without having to stack the 2-d images to vectors. Also, algorithms may benefit from not ignoring the inter-dimensional information.

For simplicity, the following sections discuss the translation invariant solution in 1-d case. Given the data $\{x_i\}$ and the corresponding label $\{y_i\}$, let f_i be the DTFT of x_i, we show the translation invariant solutions to some selected linear subspace methods and quadratic metric learning in the following sections.

The translation invariant solution to PCA. Principal component analysis (PCA), together with the linear discriminant analysis (LDA), are probably the most well-known dimension reduction and subspace methods because of their simplicity and effectiveness [10, 11]. They are often exploited as the baseline algorithms for benchmark and evaluation.

PCA tries to find a subspace whose basis vectors correspond to the maximum-variance directions in the original space. The solution to the discrete translation invariant PCA can be given in closed-form.

The translation invariant solution to LDA. Linear Discriminant Analysis (LDA) [10, 11] searches for those vectors in the underlying space that best discriminate among classes rather than those that best describe the data. More formally, given a number of independent features relative to

which the data is described, LDA creates a linear combination of these which tries to simultaneously yield the largest mean differences between the desired classes and maintain the smallest mean differences within classes. Mathematically, for all the samples of all classes, we define two measures: 1) one is called within-class scatter matrix, and 2) the other is called between-class scatter matrix.

The optimal dimension reducing transformation for LDA is the one that maximizes the between-class scatter and minimizes the within-class scatter in a reduced dimensional space. A nice discussion on these objective functions can be found in [12].

Similarly, the solution the the optimization problem of the discrete translation invariant LDA can be given in closed-form.

The translation invariant solution to DML. Xiang et al. [5] propose to learn a quadratic metric referred to as the discriminant metric learning (DML). The subtle difference between DML and LDA is, DML tries to optimize the average distance between data, while LDA tries to optimize the average distance between data and data mean; they are closely connected but not the same thing. The relationship between them is interesting but off-topic in this paper.

Due to the vastly identical form of LDA and DML, the corresponding optimization problem of translation invariant DML and its solution is given without further explanation.

We have to point out that the translation-invariant solution to DML [5] is exactly the same as the transform domain metric learning (TDML) [1].

Experiments and Discussion

In this section, experiments are performed on 3 face datasets:

1. Yale Database: Contain 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink.
2. The ORL database: Ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).
3. The extended Yale Face Database B [13]: The manually aligned and cropped images [14] are used. Contains 38 individuals and around 64 near frontal images under different illuminations per individual.

All images are aligned and cropped to 32 × 32. A random subset with p (p = 2,3,4,5,6,7,8 for Yale and ORL; p = 5,10,20,30,40,50 for YaleB) images per individual was taken with labels to form the training set, and the rest of the database was considered to be the testing set. For each given p, there are 50 randomly splits.

In the cases of PCA, LDA, DML and their translation invariant counterparts, there are two parameters to be considered. The first one is the reduction rate, which plays the same role of r. The second parameter is the sampling rate η that controls the number of samples of $g(\omega)$, e.g., $\eta \cdot p_1 p_2$ values should be sampled from the frequency spectrum for images of $p_1 \times p_2$. In practice, $1.5<\eta<2$ gives good performances. For each training process, the reduction rate is enumerated from 1/1024 to 1024/1024. For instance, 7×50×162 trainings are conducted for each method for the Yale and ORL databases, 6×50×162 for the YaleB database.

The goal of these experiments is to compare the translation invariant and non translation invariant solutions to PCA, LDA and DML. The performances are evaluated in terms of recognition rate using a nearest neighbor classifier after projecting to the transformed spaces.

Fig. 1 shows the performances of the algorithms in terms of recognition accuracy. The average recognition rate of 50 repeats for each setting is calculated and plotted. We select only one training size for each dataset (4 for Yale and ORL, 20 for YaleB), since the shape and the relative position of the curves are very similar. (All the experimental results are provided in the supplementary.) It seems a coarse conclusion can be made that the translation invariant solution usually outperforms its non translation invariant counterpart with small reduction rate. On the other hand, as reduction rate

increases, the performances of the non-TI solutions surpass the TI solutions in reverse. The overall performances can be considered as comparable. Another conclusion is that learning TI solutions in

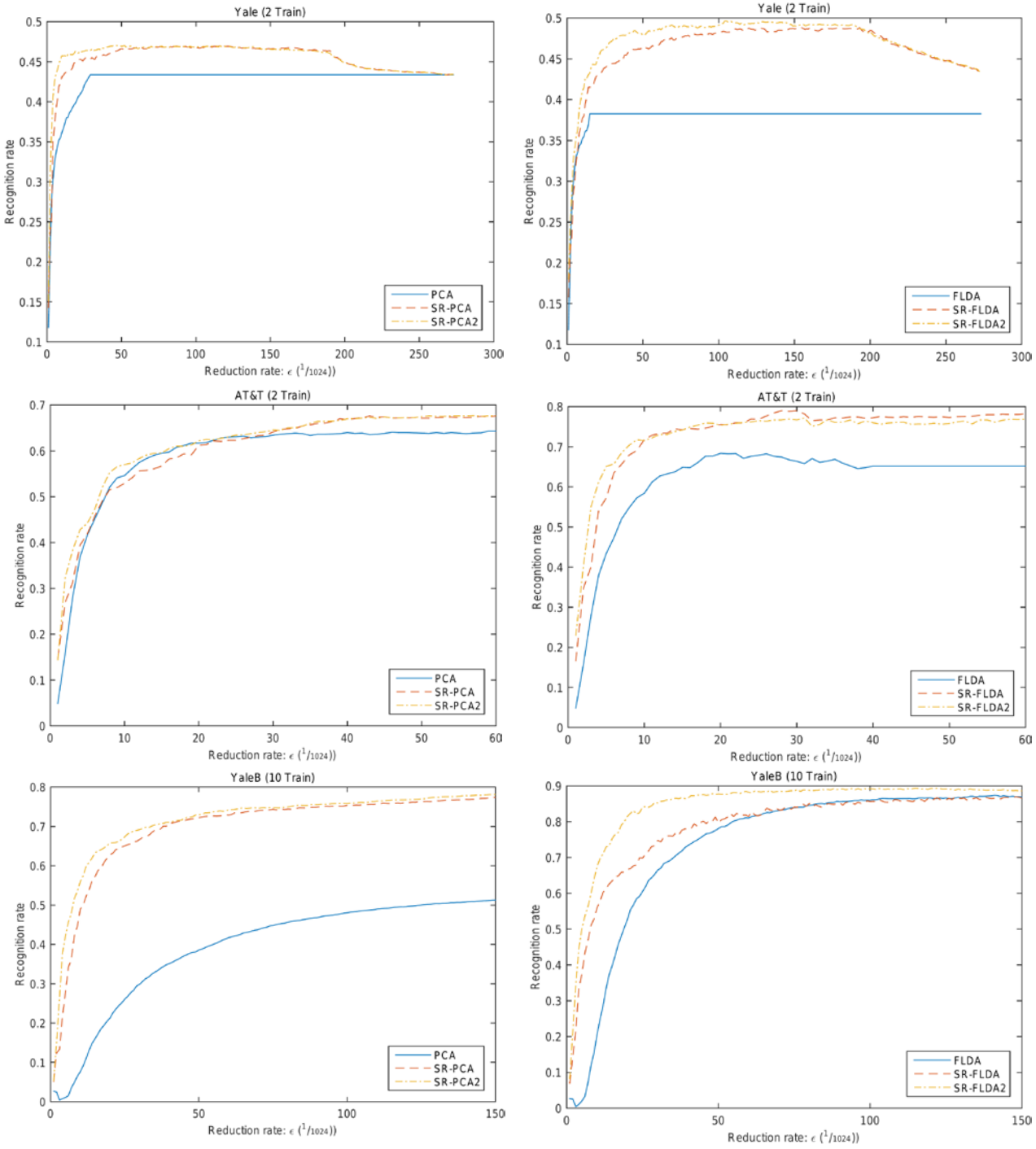

Fig. 1: The comparison of the translation invariant and non translation invariant solutions to PCA and LDA

2-d is generally better than that in 1-d, though the effect is marginal.

For completeness, part of the recognition results are also shown in Table 1. The table gives results when more training samples are provided. The recognition accuracy of TDML is also given in Table 1. It can be found that the results of TDML are identical to those of TI-LDA. This is not surprising since they are exactly the same thing when the a prior class probability is all equal.

It is worthy of mentioning the case for the YaleB database. TI-PCA behaves poor, but TI-LDA is a surprise. With reasonable small reduction rate, TI-LDA outperforms LDA by 100%~200%, which is very encouraging and interesting.

Table 1: The comparison of TI and not-TI solutions.

	Algo.	*PCA*	*TI-PCA*	*TI-PCA2*	*LDA*	*TI-LDA*	*TI-LDA2*	*TDML*	*TDML2*
0.01	Yale(2)	40.64	40.47	*47.67*	37.70	38.49	*40.33*	38.49	*40.33*
	Yale(4)	49.22	45.20	*54.76*	46.34	49.20	*54.61*	49.20	*54.61*
	Yale(8)	59.78	50.58	*65.47*	57.16	55.91	*66.31*	55.91	*66.31*
0.1	Yale(2)	45.97	48.58	*49.01*	*55.57*	49.96	50.31	49.96	50.31
	Yale(4)	54.86	55.98	*57.41*	*72.27*	57.94	58.32	57.94	58.32
	Yale(8)	63.73	65.16	*67.29*	*81.87*	68.49	68.18	68.49	68.18
0.01	ORL(2)	*60.52*	52.41	59.15	65.64	71.12	*71.58*	71.12	*71.58*
	ORL(4)	75.86	68.19	*76.94*	79.59	87.51	*87.93*	87.51	*87.93*
	ORL(8)	88.77	82.50	*91.63*	90.38	95.82	*96.53*	95.82	*96.53*
0.1	ORL(2)	70.44	70.16	*71.11*	*79.73*	72.63	72.41	72.63	72.41
	ORL(4)	83.91	84.18	*85.10*	*91.59*	85.67	85.59	85.67	85.59
	ORL(8)	93.70	94.50	*94.95*	*98.12*	94.50	94.42	94.50	94.42
0.01	YaleB(5)	*10.62*	6.66	8.96	24.46	53.31	*54.24*	53.31	*54.24*
	YaleB(20)	*16.33*	9.20	14.93	31.89	74.61	*76.08*	74.61	*76.08*
	YaleB(50)	*19.23*	10.35	18.26	*96.44*	83.17	84.32	83.17	84.32
0.1	YaleB(5)	*29.47*	19.41	22.50	*65.25*	58.93	58.21	58.93	58.21
	YaleB(20)	*53.13*	37.50	43.23	*87.18*	81.01	81.21	81.01	81.21
	YaleB(50)	*67.29*	48.65	56.06	*96.44*	89.16	89.05	89.16	89.05

Summary

In this paper, we propose the translation invariant solutions to quadratic metric learning and linear subspace methods. The translation invariant versions of optimization for PCA, LDA and DML are derived and solved. The solutions can be applied to multi-dimensional data without having to stack

them to vectors. Experiments are conducted on various benchmark datasets and the experimental results show that the TI solutions are generally on par with the non-TI solutions. Specifically, the TI solutions usually favor small sample size and small reduction dimension. The framework proposed for quadratic metric learning and linear subspace method seems quite promising.

We notice that with more training samples and large reduction dimension, the TI solutions generally suffers a performance drop. A future work is to analyze why such phenomenon emerges. Also the TI solutions are not ideal for the Yale database, hence a further analysis on the database characteristic is needed.

Acknowledgements

This work was supported by NSFC(61333015) and NBRPC(2010CB328002, 2011CB302400).

References

[1] Sun, B., Feng, J., Wang, L.: *Learning IMED via shift-invariant transformation*. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. (2009), p. 1398–1405

[2] Fukunaga, K., Flick, T.E.: *An optimal global nearest neighbor metric*. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6(3) (1984) , p. 314–318

[3] Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: *Distance metric learning, with application to clustering with side-information*. In: Advances in Neural Information Processing Systems 15. Volume 15. (2003) , p. 505–512

[4] Weinberger, K.Q., Blitzer, J., Saul, L.K.: *Distance metric learning for large margin nearest neighbor classification*. In: Advances in Neural Information Processing Systems. Volume 18. (2005) , p. 1473–1480

[5] Xiang, S., Nie, F., Zhang, C.: *Learning a mahalanobis distance metric for data clustering and classification*. Pattern Recognition 41(12) (2008) , p. 3600–3612

[6] Wang, L., Zhang, Y., Feng, J.: *On the euclidean distance of images*. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8) (2005) , p. 1334–1339

[7] Rudin, W.: *Fourier Analysis on Groups*. Wiley (January 1990)

[8] Gray, R.M.: *Toeplitz and circulant matrices: A review*. Foundations and Trends in Communications and Information Theory 2(3) (2006) , p. 155–239

[9] Lang, S.: *Algebra*. 3rd edn. Springer, Berlin (2005)

[10] Fukunaga, K.: *Introduction to Statistical Pattern Recognition*, Second Edition. Academic Press, Boston, MA (1990)

[11] Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*. 2nd edn. Wiley-Interscience (2000)

[12] Wang, H., Yan, S., Xu, D., Tang, X., Huang, T.: *Trace ratio vs. ratio trace for dimensionality reduction*. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07. (2007) , p. 1–8

[13] Georghiades, A., Belhumeur, P., Kriegman, D.: *From few to many: illumination cone models for face recognition under variable lighting and pose*. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6) (2001) , p. 643–660

[14] Lee, K.C., Ho, J., Kriegman, D.: *Acquiring linear subspaces for face recognition under variable lighting*. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(5) (2005) , p. 684–698