Optimal Metric Selection for Improved Multi-pose Face Recognition with Group Information

Xin Zhang†, Duc-Son Pham†, Wanquan Liu†, and Svetha Venkatesh⋆
†IMPCA, Curtin University, Australia and ⋆PRaDA, Deakin University, Australia
x.zhang@postgrad.curtin.edu.au, dspham@ieee.org, w.liu@curtin.edu.au, svetha.venkatesh@deakin.edu.au

Abstract

We address the limitation of sparse representation based classification with group information for multi-pose face recognition. First,we observe that the key issue of such classification problem lies in the choice of the metric norm of the residual vectors, which represent the fitness of each class. Then we point out that limitation of the current sparse representation classification algorithms is the wrong choice of the ℓ_2 norm, which does not match with data statistics as these residual values may be considerably non-Gaussian. We propose an explicit but effective solution using ℓ_p norm and explain theoretically and numerically why such metric norm would be able to suppress outliers and thus can significantly improve classification performance comparable to the state-of-arts algorithms on some challenging datasets.

1. Introduction

Recent developments in the face recognition literature have seen an increasing interest in using sparse representation [8], which is inspired by the success of compressed sensing theory. The first work in this area is [14], which proposes sparse representation classification (SRC) when dealing with extreme variations on lighting and occlusions. The SRC algorithm consists of two parts: the sparse representation via Lasso-type ℓ_1-minimization and the nearest subspace classification using ℓ_2-norm. Essentially, the algorithm represents a given face image as a sparse linear combination of other faces in the dataset, and determines which group of images corresponding to different individuals would give the best fit that determines classification.

Extension of SRC has been discussed in the literature [6, 10]. Group sparse classification (GSC) [6] extends SRC in the same way as group Lasso [15] extends the Lasso. The main argument in [6] is that the sparse solution in SRC does not favor grouping of correlated samples when Lasso regularization is used. Several modifications of GSC are also proposed in [4, 16]. We noted that the SRC was not directly compared with advanced techniques in the face recognition literature. Subsequent work has examined the performance of SRC relative to other face recognition techniques. One of which is [11] which demonstrated that sparse representation is not essential for classification at all, and that non-sparse representation performed equally well if not even better.

Given the lack of in-depth studies on the role of sparse representation in classification, there are two research questions that we will address in this work:

- What cause SRC's limitation as pointed out by recent studies and if it is possible to improve?
- Does group information really provide significant advantage for sparse representation in classification of facial images?

We will demonstrate subsequently in this paper that the main limitation of SRC is the wrong choice of ℓ_2 metric norm for the nearest subspace classification, as real residual data deviates significantly from the Gaussian and that using metric norms that can suppress heavy tails would provide a significant boost of sparse representation to the state-of-art performance. Our work is inspired by a current result in robust CS [9]. Secondly, in contrast to previous studies we found that group information does not provide significant advantage for classification as claimed previously. We will validate our findings on two benchmark datasets, the PIE and YaleB datasets.

2. Group Sparse Classification

Under the sparsity-induced classification paradigm, the first step is to express a given face image $\mathbf{x}$ as a sparse linear combination of other training images $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ so that its approximation is given by $\hat{\mathbf{x}} = \sum_{i=1}^{n} a_i \mathbf{x}_i = \mathbf{X}\mathbf{a}$ where $\mathbf{a} = [a_1 \ldots, a_n]$ is the coefficient vector. In SRC, the coefficient vector is sought to be sparse by solving

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1. \quad (1)$$

Now suppose further that the training images are naturally divided into g groups so that $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_g]$. In multi-pose face recognition, such grouping can be based on the pose information. Denote $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_g$ as the subsets of the vector coefficient $\mathbf{a}$ corresponding to those groups. Then the GSC seeks a group sparse solution via

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 + \lambda \sum_{i=1}^{g} \|\mathbf{a}_i\|_2. \quad (2)$$

We note that there are greedy formulations for both SRC and GSC, but the Lagrangian formulation described above can be solved with convex optimization algorithms that provide better numerical accuracy and stability. As can be seen, the difference between SRC and GSC is essentially the choice of the regularization term. Group sparse implies sparse, but not the converse.

The other part, which we believe more important, is the nearest subspace classification. This is accomplished by computing the fitness of each individual subspace with respect to the sparse solution. Denote as $\mathbf{X}^k$ the subset of the training images corresponding faces of class k, and $\mathbf{a}^k$ the corresponding coefficient subset. The fitness for class k is represented by the residual vector $\mathbf{r}_k = \mathbf{x} - \mathbf{X}^k\mathbf{a}^k$. In previous works, the score for such a fitness is computed via the ℓ_2-norm

$$d_k = \|\mathbf{r}_k\|_2, \quad (3)$$

and the class with a minimum score is selected.

3. Optimal Metric Selection

In (3), the score for the residual vector is calculated by ℓ_2-norm and this has been the practice without any scrutiny. In statistics, it is known that the ℓ_2 norm is optimal in the maximum likelihood sense when the residual values are approximately Gaussian. However, if there are outliers in the residual values or if the empirical distribution of the residual values has heavy tails and depart considerably from Gaussian, such ℓ_2 metric norm would be poor because it can be easily influenced by these bad outliers.

Our numerical investigation of the residual vectors from SRC reveals that it is actually the case here. In other words, we found the original SRC algorithm makes wrong decision for top candidates (with minimal scores) due to the presence of large residual values. *Where do these large values come from?* We note in the formulation (2), we optimize Gaussian-like criteria for residual from *all* classes. But in (3) we use the solution of all classes to compute the residual of each *individual* class. While the fitting term in (2) promotes Gaussian-like residual values for all classes, there is no such warranty in (3).

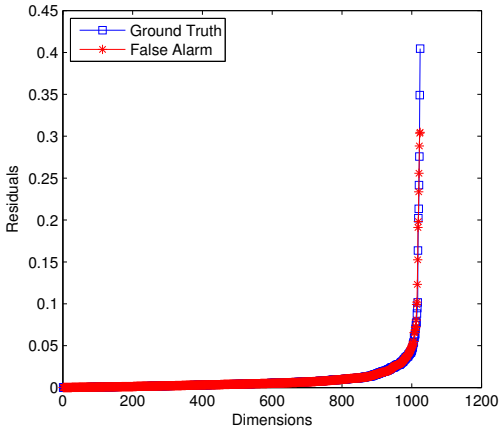

Figure 1. Residual values of top candidates.

We observe that a large number of incorrect classification decision made by SRC and GSC is when the top two candidates are very similar overall but the true candidate is tempered with some large-residual values. To illustrate, we show such a scenario in Fig. 1. We note that the residual values corresponding to the true candidates have some large values but otherwise will fit better than the incorrectly picked candidate by SRC.

The above discussion necessitates a better scheme computing the score d_k of the residual vectors $\mathbf{r}_k$ other than the ℓ_2-norm. Such a scheme must be able to detect and suppress outliers in the residual values so as to being more robust. To achieve this, one might follow robust statistics to design better score or optimize general score functions. In this work, we propose a much simple strategy by using the ℓ_p-norm, which is only controlled by one parameter, the order p of the norm.

Let us discuss why such a metric norm is useful in achieving the goal, especially when $p < 1$. Consider an oversimplified illustration in Fig. 2 where $\mathbf{r}_1 = OA$ and $\mathbf{r}_2 = OB$. Here, dimension 2 is where the outliers are present. Clearly, $\|\mathbf{r}_1\|_2 < \|\mathbf{r}_2\|_2$ as A lies in a smaller ℓ_2 ball. Suppose that dimension 1 determines the fitness then we would like to select $\mathbf{r}_2$. This is possible for some small p such that B lies on a "smaller" ℓ_p ball. Effectively, the ℓ_p ball has suppressed the outliers in dimension 2, and thus $\|\mathbf{r}_1\|_p > \|\mathbf{r}_2\|_p$. We note that making p smaller suppresses large values and effectively amplifies small values. However, making p too small may suppress too many mid values and hence would reduce the performance. In our work, we select the optimal p from the validation set.

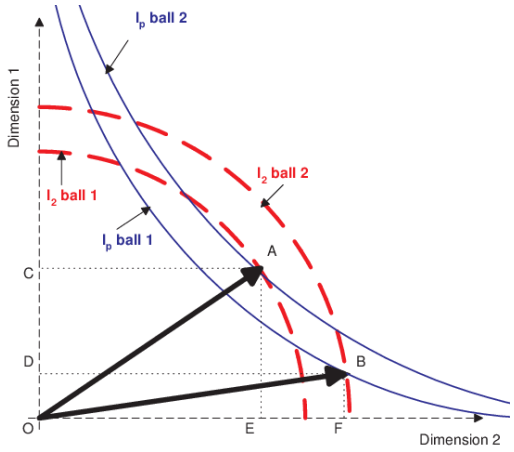

Figure 2. Why does ℓ_p norm suppress outliers?

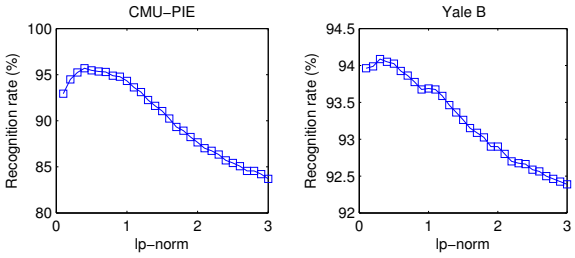

Figure 3. Recognition rates with different ℓ_p-norm

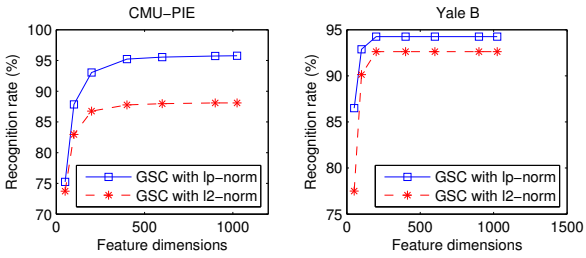

Figure 4. Dimension reduction based on PCA

4. Experiments

In this section, we present results on the widely-used CMU-PIE [12] and Yale B databases [5]. The CMU-PIE database consists of 41,368 images of 68 individuals with 13 different poses, 43 illumination conditions, and with 4 different expressions. In the Yale B database, there are 10 people with 9 different poses that combined with 64 illumination conditions for each individuals. All images are cropped and normalized to 32×32 pixels with eyes and mouth properly aligned. PCA is then applied to the centralized data to achieve group orthogonality, which improves group Lasso's numerical property.

Using random sampling, we create training, testing, and validation (for selecting optimal parameters) sets. For GSC, we use pose label information available from the datasets to create groups. We measure the performance over 10 random splits and report the average. For group Lasso, we use the advanced ADMM implementation, which is available from `http://www.stanford.edu/~boyd/papers/admm/`.

Performance dependence on ℓ_p-norm. To demonstrate how the ℓ_p-norm influences GSC's classification performance, we construct training and testing sets with 2 and 10 images per pose from each subject across all pose variations and vary the ℓ_p norm in the range between 0.1 and 3. The average classification performance is shown in Fig. 3. As can be seen, the GSC with ℓ_p norm metric in classification has the highest recognition rate at 95.8%, when p=0.5 in CMU-PIE and it achieves the highest recognition rate at p=0.3 in Yale B. Whereas, the ℓ_2 metric can not reach a satisfied rates in both CMU-PIE and Yale B. We note particularly that these plots are *interesting* as they clearly support our claim in the previous sections. These curves show two things: 1. classification with different norm metric can provide various recognition rates; 2. ℓ_2-norm metric in classification of GSC can not achieve a fair recognition rate. Thus, these observations confirms our claim that using utilized ℓ_p-norm metric to compute the residuals can improve the performance of GSC.

Performance dependence on dimensions. We next investigate how classification depends on the data's PCA dimension. Both utilized ℓ_p and ℓ_2 norm metric in GSC are tested. The p for utilized ℓ_p-norm is selected by optimizing over the validation set.

Fig. 4 shows classification performance as the dimension is varied. In both CMU-PIE and Yale B, the performance of ℓ_p-norm is always superior to ℓ_2-norm metric under any feature dimensions. When the feature length is low, the utilized ℓ_p-norm can improve about 9% in Yale B. Once the dimension length is above 300, the recognition rates of utilized ℓ_p-norm are 2% (Yale B) or 6% (CMU-PIE) higher than normal ℓ_2-norm. The highest recognition rates are achieved by ℓ_p-norm with all 1024 features in both CMU-PIE and Yale B.

In conclusion, GSC is sensitive to feature dimensions when feature length is less than 300. However, if the feature length above 300, GSC is robust under various dimensions. In addition, the experiments also show that the ℓ_p-norm can always improve the recognition rates in GSC classification under all dimensions. When the number of feature dimensions is beyond certain point, the ℓ_p-norm metric can increase the performance significantly.

Performance dependence on group number. To do

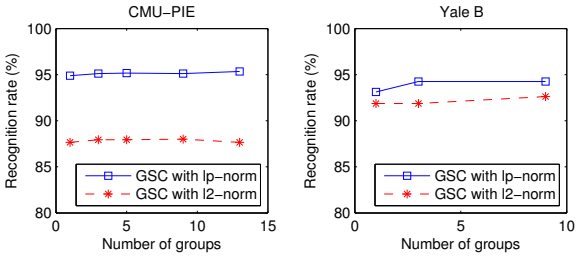

Figure 5. With various number of groups

this, we merge some groups together based on similarity of pose variations to show the effect of group information. The groups are merged by setting pose labels to the same. Both utilized GSC and normal GSC are tested in this section. The resulting curve for CMU-PIE in Figure 5 clearly shows that the recognition rates are slightly different among different number of groups. This means group information can not provide classification advantages. However, GSC with ℓ_p-norm metric can consistently improve the recognition performance with any number of groups. The findings in this experiment also are consistent with the previous experiment.

Comparison with other state-of-art algorithms. We next compare GSC with the ℓ_p-norm metric and other state-of-the-art algorithms in face recognition. Table 1 lists performance of some advanced methods and our proposed GSC. When we use the ℓ_p-norm metric in GSC classification, the performance is dramatically improved and achieves the highest recognition rate at 95.17% in CMU-PIE and 94.03% in Yale B.

Table 1. Classification Performance Comparison

Recognition rate ($\mu\% \pm \sigma\%$)		
	CMU PIE	Yale B
PCA [13]	55.17 ± 0.78	57.03 ± 1.98
LPP [7]	89.29 ± 0.63	91.59 ± 1.03
OLPP [2]	85.81 ± 0.77	92.64 ± 0.58
Regularized LDA [1]	94.88 ± 0.28	93.75 ± 0.85
Smooth LDA [3]	94.47 ± 0.24	89.90 ± 1.55
SRC [14]	89.25 ± 0.41	93.19 ± 0.22
GRC with p=2	86.84 ± 0.52	92.90 ± 1.40
GRC with utilized p	$\mathbf{95.17 \pm 0.43}$	$\mathbf{94.03 \pm 1.30}$

5. Conclusion

We have demonstrated that suitable metric norm is the key issue for improving group sparse classification, which could be simply and effectively achieved by utilizing the ℓ_p norm instead of the ℓ_2 norm. It also appears that that the group information provides little advantage in face recognition with large-pose variations. This demonstrates that future work should concentrate on further optimizing the metric norm, rather than concentrating on the group sparse representations. The extensive experiments on CMU-PIE and Yale B show that the GRC with an optimal ℓ_p-norm metric can outperform the existing state-of-the-art methods.

References

[1] D. Cai, X. He, and J. Han. SRDA: An efficient algorithm for large-scale discriminant analysis. *IEEE TKDE*, 20(1):1–12, 2008.

[2] D. Cai, X. He, J. Han, and H. Zhang. Orthogonal Laplacianfaces for face recognition. *IEEE Trans. Image. Process.*, 15(11):3608–3614, 2006.

[3] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. Learning a spatially smooth subspace for face recognition. In *Proc. CVPR*, pages 1–7. IEEE, 2007.

[4] Y. Chao, Y. Yeh, Y. Chen, Y. Lee, and Y. Wang. Locality-constrained group sparse representation for robust face recognition. In *Proc. ICIP*, pages 761–764. IEEE, 2011.

[5] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE PAMI*, 23(6):643–660, 2001.

[6] A. Majumdar and R. Ward. Classification via group sparsity promoting regularization. In *Proc. ICASSP*, pages 861–864. IEEE, 2009.

[7] X. Niyogi. Locality preserving projections. In *Proc. NIPS*, volume 16, page 153. The MIT Press, 2004.

[8] D. S. Pham and S. Venkatesh. Joint learning and dictionary construction for pattern recognition. In *Proc. CVPR*, 2008.

[9] D. S. Pham and S. Venkatesh. Improved image recovery from compressed data contaminated with impulsive noise. *IEEE Trans.Image Process.*, 21(1):397–405, 2012.

[10] H. Qiu, D. Pham, S. Venkatesh, W. Liu, and J. Lai. A fast extension for sparse representation on robust face recognition. In *Proc. ICPR*, 2010.

[11] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen. Is face recognition really a compressive sensing problem? In *Proc. CVPR*, pages 553–560. IEEE, 2011.

[12] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression (PIE) Database. In *Proc. AFGR*, May 2002.

[13] M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proc. CVPR*, pages 586–591. IEEE, 1991.

[14] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE PAMI*, 31(2):210–227, 2009.

[15] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 68(1):49–67, 2006.

[16] X. Yuan and S. Yan. Visual classification with multi-task joint sparse representation. In *Proc. CVPR*, pages 3493–3500. IEEE, 2010.