Sparsity and Robustness in Face Recognition

A tutorial on how to apply the models and tools correctly

John Wright, Arvind Ganesh, Allen Yang, Zihan Zhou, and Yi Ma

Background. This note concerns the use of techniques for sparse signal representation and sparse error correction for automatic face recognition. Much of the recent interest in these techniques comes from the paper [WYG+09], which showed how, under certain technical conditions, one could cast the face recognition problem as one of seeking a sparse representation of a given input face image in terms of a "dictionary" of training images and images of individual pixels. To be more precise, the method of [WYG+09] assumes access to a sufficient number of well-aligned training images of each of the k subjects. These images are stacked as the columns of matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$. Given a new test image $\boldsymbol{y}$, also well aligned, but possibly subject to illumination variation or occlusion, the method of [WYG+09] seeks to represent $\boldsymbol{y}$ as a sparse linear combination of the database as whole. Writing $\boldsymbol{A} = [\boldsymbol{A}_1 \mid \cdots \mid \boldsymbol{A}_k]$, this approach solves

$$\text{minimize} \quad \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subj. to} \quad \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} = \boldsymbol{y}.$$

If we let $\boldsymbol{x}_j$ denote the subvector of $\boldsymbol{x}$ corresponding to images of subject j, [WYG+09] assigns as the identity of the test image $\boldsymbol{y}$ the index whose sparse coefficients minimize the residual:

$$\hat{i} = \arg\min_i \ \|\boldsymbol{y} - \boldsymbol{A}_i\boldsymbol{x}_i - \boldsymbol{e}\|_2.$$

This approach demonstrated successful results in laboratory settings (fixed pose, varying illumination, moderate occlusion) in [WYG+09], and was extended to more realistic settings (involving moderate pose and misalignemnt) in [WWG+11]. For the sake of clarity, we repeat the above algorithm below.

$$\text{(SRC)} \begin{cases} \text{minimize} \quad \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subj. to} \quad \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} = \boldsymbol{y}, \\ \hat{i} = \arg\min_i \ \|\boldsymbol{y} - \boldsymbol{A}_i\boldsymbol{x}_i - \boldsymbol{e}\|_2. \end{cases} \tag{0.1}$$

We label this algorithm SRC (sparse representation-based classification), following the naming convention of [WYG+09].

A recent paper of Shi and collaborators [SEvdHS11] raises a number of criticisms of this approach. In particular, [SEvdHS11] suggests that (a) linear representations of the test image $\boldsymbol{y}$ in terms of training images $\boldsymbol{A}_1 \ldots \boldsymbol{A}_k$ are not well-founded and (b) that the ℓ^1-minimization in (0.1) can be replaced with a solution that minimizes the ℓ^2 residual. In this note, we briefly discuss the analytical and empirical justifications for the method of [WYG+09], as well as the implications of the criticisms of [SEvdHS11] for robust face recognition. We hope that discussing the discrepancy between the two papers within the context of a richer set of related results will provide a useful tutorial for readers who are new to these concepts and tools, helping to understand their strengths and limitations, and to apply them correctly.

1 Linear Models for Face Recognition with Varying Illumination

The method of [WYG+09] is based on low-dimensional linear models for illumination variation in face recognition. Namely, the paper assumes that if we have observed a sufficient number of well-aligned training samples $\boldsymbol{a}_1 \dots \boldsymbol{a}_n$ of a given subject j, then given a new test image $\boldsymbol{y}$ of the same subject, we can write

$$\boldsymbol{y} \approx [\boldsymbol{a}_1 \mid \cdots \mid \boldsymbol{a}_n]\, \boldsymbol{x} \doteq \boldsymbol{A}_j \boldsymbol{x}, \tag{1.1}$$

where $\boldsymbol{x}$ is a vector of coefficients. This low-dimensional linear approximation is motivated by theoretical results [BJ03, FSB04, Ram02] showing that well-aligned images of a convex, Lambertian object lie near a low-dimensional linear subspace of the high-dimensional image space. These results were themselves motivated by a wealth of previous empirical evidence of effectiveness of linear subspace approximations for illumination variation in face data (see [Hal94, EHY95, BK98, YSEB99, GBK01]).

To see this phenomenon in the data used in [WYG+09], we take Subsets 1-3 of the Extended Yale B database (as used in the experiments by [WYG+09]). We compute the singular value decomposition of each subject's images. Figure 1 (left) plots the mean of each singular value, across all 38 subjects. We observe that most of the energy is concentrated in the first few singular values.

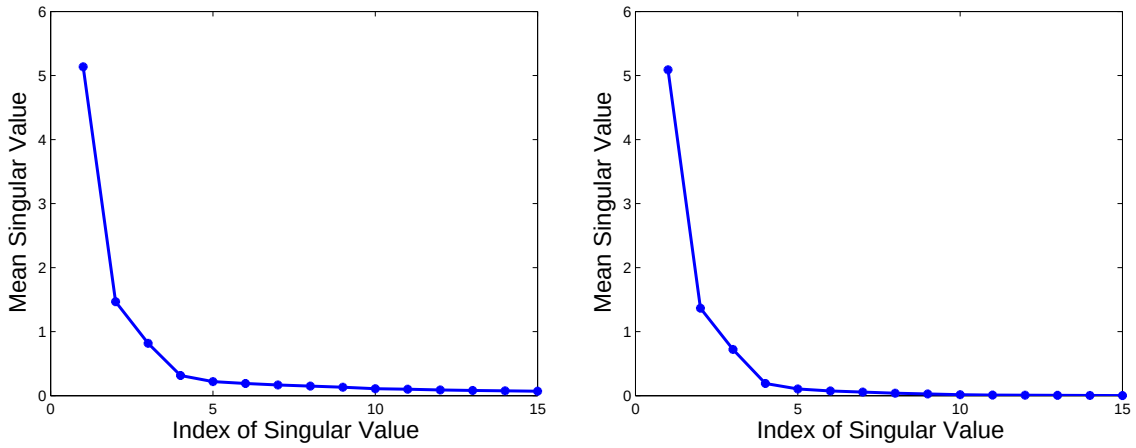
Of course, some care is necessary in using these observations to construct algorithms. The following physical phenomena break the low-dimensional linear model:

- **Specularities and cast shadows** break the assumptions of the low-dimensional linear model. These phenomena are spatially localized, and can be treated as large-magnitude, sparse errors.
- **Occlusion** also introduces large-magnitude, sparse errors.
- **Pose variations and misalignment** introduce highly nonlinear transformations of domain, which break the low-dimensional linear model.

Specularities, cast shadows and moderate occlusion can be handled using techniques from sparse error correction. Indeed, using the "Robust PCA" technique of [CLMW11] to remove sparse errors due to cast shadows and specularities, we obtain Figure 1 (right). Once violations of the linear model are corrected, the singular values decay more quickly. Indeed, only the first 9 singular values are significant, corroborating theoretical results of Basri, Ramamoorthi and collaborators.

The work of [WYG+09] assumed access to well-aligned training images, with sufficient illuminations to accurately approximate new input images. Whether this assumption holds in practice depends strongly on the scenario. In extreme examples, when only a single training image per subject is available, it will clearly be violated. In applications to security and access control, this assumption can be met: [WWG+11] discusses how to collect sufficient training data for a single subject, and how to deal with misalignment in the test image. Less controlled training data (for example, subject to misalignment) can be dealt with using similar techniques [PGX+11].

The above experiments use the Extended Yale B face database, which was constructed to investigate illumination variations in face recognition. However, similar results can be obtained on other datasets. We demonstrate this using the AR database, which was also used in the experiments of [WYG+09]. We take the cropped images from this database, with varying expression and illumination. There are a total of 14 images per subject. Figure 2 plots the resulting singular values obtained



Low rank approximation by SVD **Low-rank approximation by RPCA**

Figure 1: **Low-dimesional structure in the Extended Yale B database.** We compute low-rank approximations to the images of each subject in the Extended Yale B database, under illumination subsets 1-3. (left) Mean singular values across subjects, when low-rank approximation is computed using singular value decomposition. (right) Mean singular values across subjects, when low-rank approximation is computed robustly using convex optimization. In both cases, the singular values decay; when sparse errors are corrected, the decay is more pronounced.

via SVD (left) and with a robust low-rank approximation (right). One can clearly observe low-rank structure[1]. However, this structure does not necessarily arise from the Lambertian model – the number of distinct illuminations may not be sufficient, and some subjects' images have significant saturation. Rather, the low-rank structure in the AR database arises from the fact that conditions are repeated over time.

Comments on the "assumption test" by Shi et. al. [SEvdHS11] report the following experimental result: all of the cropped images from *all subjects* of the AR database are stacked as columns of a large matrix $\boldsymbol{A}$. The singular values of $\boldsymbol{A}$ are computed. The singular values of this matrix are peaked in the first few entries, but have a heavy tail. Because of this, [SEvdHS11] conclude that images of *a single subject* in AR do not exhibit low-dimensional linear structure. Their observation does not imply this conclusion, for at least two reasons:

- First, low-dimensional linear structure is expected to occur within the images of a single subject. The distribution of singular values of a dataset of many subjects as a whole depends not only on the physical properties of each subject's images, but on the distribution of face shapes and reflectances across the population of interest. Investigating properties of the singular values of the database as a whole is a questionable way to test hypotheses about the numerical rank or spectrum of a single subject's images. This is especially the case when each subject's

[1]In fact, when the low-rank approximation is computed robustly, its numerical rank always lies in the range of $6-9$. However, this number is less important than the singular values themselves, which decay quickly.

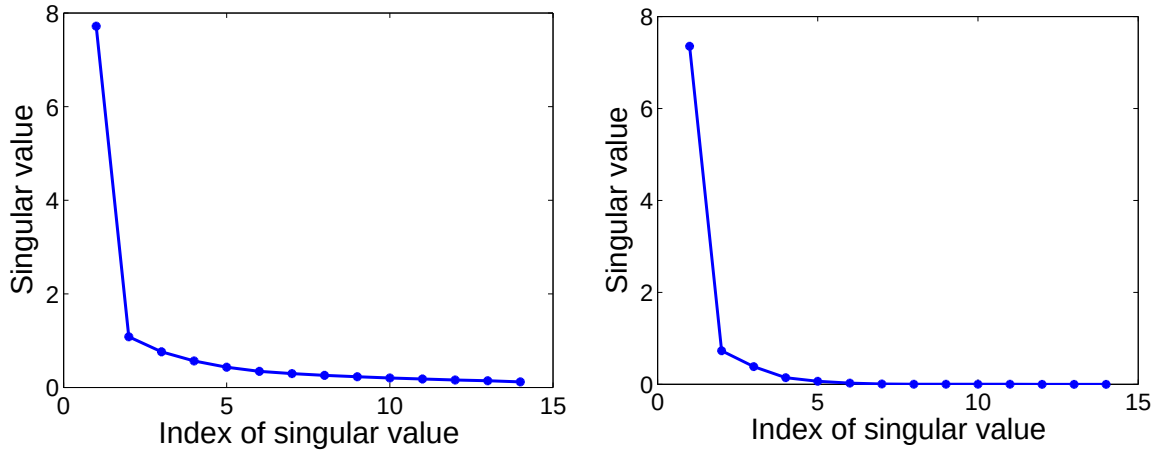


Figure 2: **Low-dimensional structure in the AR database.** We compute low-rank approximations to the images of each subject in the AR database, using images with varying illumination and expression (14 images per subject). (left) Mean singular values across subjects, when low-rank approximation is computed using singular value decomposition. (right) Mean singular values across subjects, when low-rank approximation is computed robustly using convex optimization. Again, in both cases, the singular values decay; when sparse errors are corrected, the decay is more pronounced.

images are not perfectly rank deficient, but rather approximated by a low-dimensional subspace (as is implied by [BJ03]): the overall spectrum of the matrix will depend significantly on the relative orientation of all the subspaces.[2]

- Second, the images used in the experiment of Shi et. al. include occlusions, and may not be precisely aligned at the pixel level. Both of these effects are known to break low-dimensional linear models. Indeed, above, we saw that if we restrict our attention to training images that do not have occlusion (as in [WYG+09]) and compute robustly, low-dimensional linear structure becomes evident.

2 Robustness, ℓ^1 and the ℓ^2 Alternatives

In the previous section, we saw that images of the same face under varying illumination could be well-represented using a low-dimensional linear subspace, provided they were well-aligned and provided one could correct gross errors due to cast shadows and specularities. These errors are

[2]Indeed, [SEvdHS11] observe a distribution of singular values across all the subjects that resembles the singular values of a Gaussian matrix. This is reminiscent of [WM10], in which the the uncorrupted training images of many subjects are modeled as small Gaussian deviations about a common mean. The implications of such a model for error correction are rigorously analyzed in [WM10]. It should also be noted that the values of the plotted singular values in [SEvdHS11] are not, as suggested, the singular values of a standard Gaussian matrix of the same size as the test database – they are the singular values of a smaller, *square* Gaussian random matrix, and hence do not reflect the noise floor in the AR database.

prevalent in real face images, as are additional violations of the linear model due to occlusion. Like specular highlights, the error incurred by occlusion can be large in magnitude, but is confined to only a fraction of the image pixels – it is *sparse* in the pixel domain. In [WYG+09], this effect is modeled using an additive error $\boldsymbol{e}$. If the only prior information we have about $\boldsymbol{e}$ is that it is sparse, then the appropriate optimization problem becomes

$$\text{minimize} \quad \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subj. to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}. \tag{2.1}$$

Clearly, any robustness derived from the solution to this optimization problem is due to the presence of the sparse error term, and the minimization of the ℓ^1 norm of $\boldsymbol{e}$. Indeed, based on theoretical results in sparse error correction, we should expect that the above ℓ^1 minimization problem will successfully correct the errors $\boldsymbol{e}$ provided the number of errors (corrupted, occluded or specular pixels) is not too large. For certain classes of matrices $\boldsymbol{A}$ one can identify sharp thresholds on the number of errors, below which ℓ^1 minimization performs perfectly, and beyond which it breaks down. In contrast, minimization of the ℓ^2 residual, say $\min \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2$ does not have this property.

The paper of [SEvdHS11] suggests that the use of the ℓ^1 norm in (2.1) is unnecessary, and proposes two algorithms. The first solves

$$(\ell^2\text{-}1) \quad \begin{cases} \text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2, \\ \hat{i} = \arg\min_i \ \|\boldsymbol{y} - \boldsymbol{A}_i\boldsymbol{x}_i\|_2. \end{cases} \tag{2.2}$$

This approach is not expected to be robust to errors or occlusion. For faces occluded with sunglasses and scarves (as in the AR Face Database), [SEvdHS11] suggests an extension

$$(\ell^2\text{-}2) \quad \begin{cases} \text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{W}\boldsymbol{v}\|_2, \\ \hat{i} = \arg\min_i \ \|\boldsymbol{y} - \boldsymbol{A}_i\boldsymbol{x}_i\|_2. \end{cases} \tag{2.3}$$

where $\boldsymbol{W}$ is a tall matrix whose columns are chosen as blocks that may well-represent occlusions of this nature.

In trying to understand the strengths and working conditions of these proposals several questions arise. First, do the approaches (SRC), (ℓ^2-1) and (ℓ^2-2) provide robustness to general pixel-sparse errors? We test this using settings and data *identical* to those in [WYG+09], in which the Extended Yale B database subsets I and II are used for training, and subset III is used for testing. Varying fractions of random pixel corruption are added, from 0% to 90%. Table 1 shows the resulting recognition rates for the three algorithms. The ℓ^1 minimization (2.1) is robust to up to 60-70% arbitrary random errors. In contrast, both methods based on ℓ^2 minimization break down much more quickly. We note that this result is expected from theory: [WM10] provides results in this direction.[3] To be clear, the goal of this experiment is not to assert that the ℓ^1 norm is "better" or "worse" than ℓ^2 in some general sense – simply to show that ℓ^1 provides robustness to general sparse errors, whereas the two approaches (2.2)-(2.3) do not. There are situations in which it is correct (optimal, in fact) to minimize the ℓ^2 norm – when the error is expected to be dense, and in particular, if it follows an iid Gaussian prior. However, for sparse errors, ℓ^1 has well-justified and thoroughly documented advantages.

Of course, real occlusions in images are very different in nature for the random corruptions considered above – occlusions are often spatially contiguous, for example. Hence, we next ask to

[3] To be precise, results in [WM10] suggest, but do not prove, that ℓ^1 will succeed at correcting large fractions of errors in this situation. The rigorous theoretical results of [WM10] pertain to a specific stochastic model for $\boldsymbol{A}$.

% corrupted pixels	Recognition rate (%) SRC	ℓ^2-1	ℓ^2-2
0	100	100	100
10	100	100	100
20	**100**	99.78	99.78
30	**100**	99.56	99.34
40	**100**	96.25	96.03
50	**100**	83.44	81.23
60	**99.3**	59.38	59.94
70	**90.7**	38.85	40.18
80	**37.5**	15.89	15.23
90	7.1	8.17	**7.28**

Table 1: **Extended Yale B database with random corruption.** Subsets 1 and 2 are used as training and Subset 3 as testing. The best recognition rates are in bold face. SRC (ℓ^1) performs robustly up to about 60% corruption, and then breaks down. Alternatives are significantly less robust.

what extent the three methods provide robustness against general spatially contiguous errors. We investigate this using random synthetic block occlusions *exactly the same* as in [WYG+09]. The results are reported in Table 2.

% occluded pixels	Recognition rate (%) SRC	ℓ^2-1	ℓ^2-2
10	**100**	99.56	99.78
20	**99.8**	95.36	97.79
30	**98.5**	87.42	92.72
40	**90.3**	76.82	82.56
50	65.3	60.93	**66.22**

Table 2: **Extended Yale B with block occlusions.** Subsets 1 and 2 are used as training, Subset 3 as testing. The best recognition rates are in bold face. SRC ℓ^1 minimization performs quite well upto a breakdown point near 30% occluded pixels, then breaks down. The two alternatives based on ℓ^2 norm minimization degrade more rapidly as the frraction of occlusion increases.

Notice that again, ℓ^1 minimization performs more robustly than either of the ℓ^2 alternatives. As in the previous experiment, the good performance compared to (ℓ^2-1) is expected (indeed, [SEvdHS11] do not assert that (ℓ^2-1) is robust against error). The good performance compared to (ℓ^2-2) is also expected, as the basis $\boldsymbol{W}$ is designed for certain specific errors (incurred by sunglasses and scarves). It is also important to note that the breakdown point for ℓ^1 with spatially coherent errors is lower than for random errors ($\approx 30\%$ compared to $\approx 60\%$). Again, this is expected – the theory of ℓ^1 minimization suggests the existence of a worst case breakdown point (the strong threshold), which is lower than the breakdown point for randomly supported solutions (the weak threshold). For spatially coherent errors, we should not expect ℓ^1 minimization to succeed beyond this threshold of 30%. Nevertheless, if one could incorporate the spatial continuity prior of the error

support in a principled manner, one could expect to see ℓ^1 minimization to tolerate more than 60% errors, as investigated further in [ZWM$^+$09], well before the work of [SEvdHS11].

Finally, to what extent do the three methods provide robustness to the specific real occlusions encountered in the AR database? Here, we should distinguish between two cases – occlusion by sunglasses and occlusion by scarves. Sunglasses fall closer to the aforementioned threhold, whereas scarves significantly violate it, covering over 40% of the face. Table 3 shows the results of the three methods for these types of occlusion, at the same image resolution used in [WYG$^+$09] (80×60).[4]

	Recognition rate (%)			
Occlusion type	**SRC**	**ℓ^2-1**	**ℓ^2-2**	[ZWM$^+$09]
Sunglasses	87	59.5	83	99–**100**
Scarf	59.5	85	82.5	97–**97.5**

Table 3: **AR database, with the data and settings of [WYG$^+$09].** SRC outperforms ℓ^2 alternatives for sunglasses, but does not handle occlusion by scarves well, as it falls beyond the breakdown point for contiguous occlusion.

From Table 3, one can see that none of the three methods is particularly satisfactory in its performance. For sunglasses, ℓ^1 norm minimization outperforms both ℓ^2 alternatives. Scarves fall beyond the breakdown point of ℓ^1 minimization, and SRC's performance is, as expected, unsatisfactory. The performance of (ℓ^2-2) for this case is better, although none of the methods offers the strong robustness that we saw above for the Yale dataset. This is the case despite the fact that the basis $\boldsymbol{W}$ in (ℓ^2-2) was chosen specifically for real occlusions.

There may be several reasons for the above unsatisfactory results on the AR database: 1. Unlike the Yale database, the AR database does not have many illuminations and images are not particularly well aligned either – all may compromise the validity of the linear model assumed. 2. None of the models and solutions is particularly effective in exploiting the spatial continuity of the large error supports like sunglasses or scarfs.

A much more effective way of harnessing the spatial continuity of the error supports was investigated in [ZWM$^+$09], where ℓ^1 minimization, together with a Markov random field model for the errors, can achieve nearly 100% recognition rates for sunglasses and scarfs with exactly the same setting (trainings, resolution) as above experiments on the AR database.

3 Comparison on the AR Database with Full-Resolution Images

Readers versed in the literature on error correction (or ℓ^1-minimization) will recognize that its good performance is largely a *high-dimensional* phenomenon. In the previous examples, it is natural to wonder what lost when we run the methods at lower resolution (80×60). In this section, we compare the three methods at the native resolution 165×120 of the cropped AR database. This is possible thanks to scalable methods for ℓ^1 minimization [YGZ$^+$11].

We use a training set consisting of 5 images per subject – four neutral expressions under different lighting, and one anger expression, which is close to neutral, all taken under with the same expression.

[4]The basis images used in forming the matrix $\boldsymbol{W}$ are transformed to this size using Matlab's imresize command.

From the training set of [WYG+09], we removed three images with large expression (smile and scream), as these effects violate the low-dimensional linear model. In the cropped AR database, for each person, the training set consists of images 1, 3, 5, 6 and 7. The other 8 images per person from Session 1 were used for testing. Table 4 lists the recognition rates for each category of test image. Note that there are 100 test images (1 per person) in each category. For these experiments, we use an Augmented Lagrange Multiplier (ALM) algorithm to solve the ℓ^1 minimization problem (see [YGZ+11] for more details). Our Matlab implementation requires on average 259 seconds per test image, when run on a MacPro with two 2.66 GHz Dual-Core Intel Xenon processers and 4GB of memory.[5] We would like to point out that there is scope for improvement in the speed of our implementation. But since this is not the focus of our discussion here, we have used a simple, straightforward version of the ALM algorithm that is accurate but not necessarily very efficient. In addition, we have used a single-core implementation. The ALM algorithm is very easily amenable to parallelization, and this could greatly reduce the running time, especially when we have a large number of subjects in the database.

	Recognition rate (%)		
Test Image category	**SRC**	**ℓ^2-1**	**ℓ^2-2**
Smile	**100**	97	95
Scream	**88**	60	59
Sunglass (neutral lighting)	**88**	68	**88**
Sunglass (lighting 1)	75	63	**88**
Sunglass (lighting 2)	**90**	69	84
Scarf (neutral lighting)	65	66	**76**
Scarf (lighting 1)	**66**	63	65
Scarf (lighting 2)	**68**	62	67
Overall	**80**	68.5	77.75

Table 4: **AR database with 5 training images per person and full resolution**. The best recognition rates are in bold face.

From the above experiment, we can see that when the three approaches are compared with images of the same resolution, the results differ significantly from those of [SEvdHS11]. We will explain this discrepency in the next section.

On the other hand, we observe that none of the methods performs in a completely satisfactory manner on images with large occlusion – in particular, images with scarves. This is expected from our experiments in the previous section. Can strong robustness (like that exhibited by SRC with $\leq 60\%$ random errors or $\leq 30\%$ contiguous errors) be achieved here? It certainly seems plausible, since neither SRC nor (ℓ^2-1) take advantage of spatial coherence of real occlusions. (ℓ^2-2) does take advantage of spatial properties of real occlusions, through the construction of the matrix $\boldsymbol{W}$, but it is not clear if or how one can construct a $\boldsymbol{W}$ that is guaranteed to work for all practical cases.

In [WYG+09], ℓ^1-norm minimization together with a partitioning heuristic is shown to produce much improved recognition rates on the particular cases encountered in AR (97.5% for sunglasses and 93.5% for scarfs). However, the choice of partition is somewhat arbitrary, and this heuristic suffers from many of the same conceptual drawbacks as the introduction of a specific basis $\boldsymbol{W}$.

[5] With 8 images per subject, as in [WYG+09], this same approach requires 378 seconds per test image.

Several groups have studied more principled schemes for exploiting prior information on the spatial layout of sparse signals or errors (see [ZWM+09] and the references therein). For instance, one could expect that the modified ℓ^1 minimization method given in [ZWM+09] would work equally well under the setting (training and resolution) of the above experiments as it did under the setting in the previous section (see Table 3).

4 Face recognition with low-dimensional measurements

The results in the previous section, and conclusions that one may draw from them, are quite different from those obtained by Shi et. al. [SEvdHS11]. The reasons for this discrepancy are simple:

- In [SEvdHS11], the authors did not solve (0.1) to compare with [WYG+09]. Rather, they solved[6]

$$\text{minimize} \quad \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subj. to} \quad \boldsymbol{\Phi y} = \boldsymbol{\Phi}(\boldsymbol{Ax} + \boldsymbol{e}), \tag{4.1}$$

  where $\boldsymbol{\Phi}$ is a random projection matrix mapping from the $165 \times 120 = 19,800$-dimensional image space into a meager 300-dimensional feature space. Using these drastically lower (300) dimensional features, they obtain recognition rates of around 40% for the above ℓ^1 minimization, which is compared to a 78% recognition rate obtained with (ℓ^2-2) on the full $(19,800)$ image dimension. As we saw in the previous section, when the two methods are compared on a fair footing with the same number of observation dimensions, the conclusions become very different.

- In Section 5 of [SEvdHS11], there is an additional issue: the training images in $\boldsymbol{A}$ are randomly selected from the AR dataset sessions regardless of their nature. In particular, the training and test sets could contain images with significant occlusion. This choice is very different from any of the experimental settings in [WYG+09],[7] and also different from settings of all of the above experiments. In Section 1, we have already discussed the problems with such a choice and how it differs from the work of [WYG+09].

The main methodological flaw of [SEvdHS11] is to compare the performance of the two methods with dramatically different numbers of measurements – and in a situation that is quite different from what was advocated in [WYG+09]:

- It is easy to see that the minimizer in (4.1) can have at most $d = 300$ nonzero entries – far less than the cardinality of the occlusion such as sun glasses or scarf. ℓ^1 minimization will not succeed in this scenario. In fact, both (ℓ^2-1) and (ℓ^2-2) also fail when applied with this set of $d = 300$ features. Without proper regularization on $\boldsymbol{x}$ (say via the ℓ^1-norm), (ℓ^2-1) and (ℓ^2-2) have infinite many minimizers, and the approach suggested in [SEvdHS11] cannot apply.

- [WYG+09] also investigated empirically the use random projections as features, for images that are not occluded or corrupted! The model is strictly $\boldsymbol{y} = \boldsymbol{Ax}$ (or $\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{z}$, where

[6] It seems likely that the authors of [SEvdHS11] mistakenly solved instead the following problem: minimize $\|\boldsymbol{x}\|_1 + \|\boldsymbol{e}'\|_1$ subj. to $\boldsymbol{\Phi y} = \boldsymbol{\Phi Ax} + \boldsymbol{e}'$. If that was the case, their results would be even more problematic as the projected error $\boldsymbol{e}' = \Phi\boldsymbol{e}$ is no longer sparse for an arbitrary random projection. In practice, the sparsity of $\boldsymbol{e}'$ can only be ensured if the projection is a simple downsampling.

[7] In [SEvdHS11], the authors claim that they "form the matrix $\boldsymbol{A}$ in the same manner as [WYG+09]". That is simply *not true*.

$\boldsymbol{z}$ is small (Gaussian) noise) – no gross errors are involved. As the problem of solving for $\boldsymbol{x}$ from $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ is underdetermined, ℓ^1 regularization on $\boldsymbol{x}$ becomes necessarily to obtain the correct solution. However, [WYG+09] does not suggest that a random projection into a lower-dimensional space can improve robustness – this is provably false. It also does not suggest solving (4.1) in cases with errors – as the results of [SEvdHS11] suggest, this does not work particularly well.

- Nevertheless, *under very special conditions*, robustness can still be achieved with severely low-dimensional measurements. As investigated in [ZWM+09], if the low-dimensional measures are from down-sampling (that respects the spatial continuity of the errors) and the spatial continuity of the error supports is effectively exploited using a Markov random field model, one can achieve nearly 90% recognition rates for scarfs and sunglasses at the resolution of 13×9 – only 111 measurements (pixels), far below the 300 (random) measurements used in [SEvdHS11].

5 Linear models and solutions

Like face recognition, many other problems in computer vision or pattern recognition can be cast as solving a set of linear equations, $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$. Some care is necessary to do this correctly:

1. The first step is to verify that the linear model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ is valid, ideally via physical modeling corroborated by numerical experiments. If the training $\boldsymbol{A}$ and the test $\boldsymbol{y}$ are not prepared in a way such a model is valid, two things could happen: 1. there might be no solution or no (unique) solution to the equations; 2. the solution can be irrelevant to what you want.

2. The second step, based on the properties of the desired $\boldsymbol{x}$ (least energy or entropy) and those of the errors $\boldsymbol{e}$ (dense Gaussian or sparse Laplacian), one needs to choose the correct optimization objective in order to obtain the correct solution.

There are already four possible combinations of ℓ^1 and ℓ^2 norms[8]:

$$\begin{array}{llll}
\text{minimize} & \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 & \text{subj. to} & \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} \quad \text{(least entropy \& error correction)} \\
\text{minimize} & \|\boldsymbol{x}\|_2 + \|\boldsymbol{e}\|_1 & \text{subj. to} & \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} \quad \text{(least energy \& error correction)} \\
\text{minimize} & \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_2 & \text{subj. to} & \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} \quad \text{(sparse regression with noise – lasso)} \\
\text{minimize} & \|\boldsymbol{x}\|_2 + \|\boldsymbol{e}\|_2 & \text{subj. to} & \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} \quad \text{(least energy with noise)}
\end{array}$$

Ideally, the question should not be which formulation yields better performance on a specific dataset, but rather which assumptions match the setting of the problem, and then whether the adopted regularizer helps find the correct solution under these assumptions. For instance, when $\boldsymbol{A}$ is under-determined, regularization on $\boldsymbol{x}$ with either the ℓ^1 or the ℓ^2 norm is necessary to ensure a unique solution. But the solution can be rather different for each norm. If $\boldsymbol{A}$ is over-determined, the choice of regularizer on $\boldsymbol{x}$ is less important or even is unnecessary. Furthermore, be aware that all above programs could fail (to find the correct solution) beyond their range of working conditions. Beyond the range, it becomes necessary to exploit additional structure or information about the signals ($\boldsymbol{x}$ or $\boldsymbol{e}$) such as spatial continuity etc.

[8] In the literature, many other norms are also being investigated such as the $\ell^{2,1}$ norm for block sparsity etc.

References

[BJ03] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

[BK98] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal on Computer Vision*, 28(3):245–260, 1998.

[CLMW11] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 2011.

[EHY95] R. Epstein, P. Hallinan, and A. Yuille. 5 plus or minus 2 eigenimages suffice: An empirical investigation of low-dimensional lighting models. In *IEEE Workshop on Physics-based Modeling in Computer Vision*, 1995.

[FSB04] D. Frolova, D. Simakov, and R. Basri. Accuracy of spherical harmonic approximations for images of Lambertian objects under far and near lighting. In *Proceedings of the European Conference on Computer Vision*, pages 574–587, 2004.

[GBK01] A. Georghiades, P. Belhumeur, and D. Kriegman. 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):643–660, 2001.

[Hal94] P. Hallinan. A low-dimensional representation of human faces for arbitrary lighting conditions. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 1994.

[PGX+11] Y. Peng, A. Ganesh, W. Xu, J. Wright, and Y. Ma. RASL: Robust alignment via sparse and low-rank decomposition for linearly correlated images. *preprint*, 2011.

[Ram02] R. Ramamoorthi. Analytic PCA construction for theoretical analysis of lighting variability, including attached shadows, in a single image of a convex lambertian object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[SEvdHS11] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen. Is face recognition really a compressive sensing problem? In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[WM10] J. Wright and Y. Ma. Dense error correction via ℓ^1-minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010.

[WWG+11] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma. Towards a practical automatic face recognition system: robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[WYG+09] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210 – 227, 2009.

[YGZ+11] A. Yang, A. Ganesh, Z. Zhou, S. Sastry, and Y. Ma. Fast ℓ_1-minimization algorithms for robust face recognition. *(preprint) arXiv:1007.3753*, 2011.

[YSEB99] A. Yuille, D. Snow, R. Epstein, and P. Belhumeur. Determining generative modesl of objects under varying illumination: Shape and albedo from multiple images usign SVD and integrability. *International Journal on Computer Vision*, 35(3):203–222, 1999.

[ZWM+09] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma. Face recognition with contiguous occlusion using markov random fields. In *Proceedings of the IEEE International Conference on Computer Vision*, 2009.