Unifying Nuclear Norm and Bilinear Factorization Approaches for Low-rank Matrix Decomposition

Ricardo Cabral[†,‡], Fernando De la Torre[‡], João P. Costeira[†], Alexandre Bernardino[†]
[†]ISR - Instituto Superior Técnico, Lisboa, Portugal
[‡]Carnegie Mellon University, Pittsburgh, PA, USA

rscabral@cmu.edu, ftorre@cs.cmu.edu, {jpc,alex}@isr.ist.utl.pt

Abstract

Low rank models have been widely used for the representation of shape, appearance or motion in computer vision problems. Traditional approaches to fit low rank models make use of an explicit bilinear factorization. These approaches benefit from fast numerical methods for optimization and easy kernelization. However, they suffer from serious local minima problems depending on the loss function and the amount/type of missing data. Recently, these low-rank models have alternatively been formulated as convex problems using the nuclear norm regularizer; unlike factorization methods, their numerical solvers are slow and it is unclear how to kernelize them or to impose a rank a priori.

This paper proposes a unified approach to bilinear factorization and nuclear norm regularization, that inherits the benefits of both. We analyze the conditions under which these approaches are equivalent. Moreover, based on this analysis, we propose a new optimization algorithm and a "rank continuation" strategy that outperform state-of-the-art approaches for Robust PCA, Structure from Motion and Photometric Stereo with outliers and missing data.

1. Introduction

Many computer vision, signal processing and statistical problems can be posed as problems of learning low dimensional models from data. Low rank models have been widely used for learning representations of shape, appearance or motion in computer vision problems, under several noise assumptions and use of prior information [12, 14, 15, 35]. All these problems are directly or indirectly related to the problem of recovering a rank-k matrix $\mathbf{Z}$ from a corrupted data matrix $\mathbf{X}$, by minimizing

$$\begin{aligned} \min_{\mathbf{Z}} \quad & f(\mathbf{X}-\mathbf{Z}) \\ \text{subject to} \quad & \text{rank}(\mathbf{Z}) = k, \end{aligned} \tag{1}$$

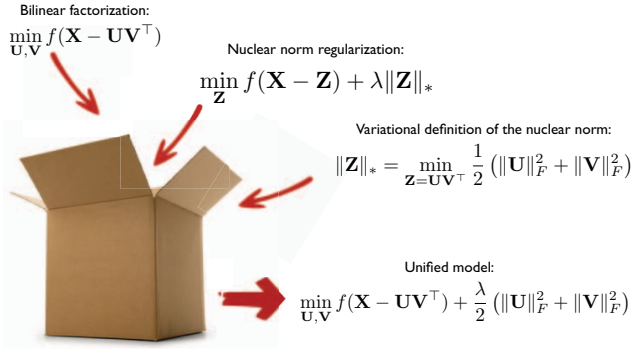


Figure 1. Low-rank matrix decomposition can be achieved with both bilinear factorization and nuclear norm regularization models. We analyze the conditions under which these are equivalent and propose a unified model that inherits the benefits of both.

where $f(\cdot)$ denotes a loss function (see footnote[1] for notation). Due to its intractability, the rank constraint in (1) has typically been imposed by a factorization $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$, as

$$\min_{\mathbf{U},\mathbf{V}} \quad f(\mathbf{X}-\mathbf{U}\mathbf{V}^\top). \tag{2}$$

It has been shown that when the loss function is the Least Squares (LS) loss, *i.e.*, $f(\mathbf{X}-\mathbf{U}\mathbf{V}^\top) = \|\mathbf{X}-\mathbf{U}\mathbf{V}^\top\|_F^2$, then (2) does not have local minima and a closed form solution can be obtained via the Singular Value Decomposition (SVD) of $\mathbf{X}$ [3]. Unfortunately, this factorization approach has several caveats: The LS loss is highly susceptible to outliers; also, the presence of missing data in $\mathbf{X}$ results in local minima. While outliers can be addressed with robust loss functions [14, 22], factorization with missing data is an

[1]Bold capital letters denote matrices (*e.g.*, $\mathbf{D}$). All non-bold letters denote scalar variables. d_{ij} denotes the scalar in the row i and column j of $\mathbf{D}$. $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ denotes the inner product between two vectors $\mathbf{d}_1$ and $\mathbf{d}_2$. $||\mathbf{d}||_2^2 = \langle \mathbf{d}, \mathbf{d} \rangle = \sum_i d_i^2$ denotes the squared Euclidean Norm of the vector $\mathbf{d}$. $\text{tr}(\mathbf{A}) = \sum_i a_{ii}$ is the trace of $\mathbf{A}$. $\|\mathbf{A}\|_F^2 = \text{tr}(\mathbf{A}^\top\mathbf{A}) = \sum_{ij} a_{ij}^2$ designates the squared Frobenius Norm of $\mathbf{A}$. $||\mathbf{A}||_*$ designates the nuclear norm (sum of singular values) of $\mathbf{A}$. $\odot$ denotes the Hadamard or element-wise product. $\mathbf{I}_K \in \mathbb{R}^{K\times K}$ denotes the identity matrix.

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.309

NP-Hard problem [17]. For this reason, the optimization of (1) remains an active research topic [5, 15, 36, 41].

Recently, several works have surfaced which exploit low-rank structure through regularization, by minimizing

$$\min_{\mathbf{Z}} \quad f(\mathbf{X}-\mathbf{Z}) + \lambda \|\mathbf{Z}\|_*, \tag{3}$$

where λ is a trade-off parameter between the loss function and the low-rank regularization induced by the nuclear norm. These models have extended the use of low-rank priors to many applications where $\mathbf{Z}$ is low rank but its rank is not known *a priori* [9, 20, 40]. Despite their convexity and theoretical guidelines for the choice of λ [9], these models also have several drawbacks. First, it is unclear how to impose a certain rank in $\mathbf{Z}$: we show that adjusting λ so $\mathbf{Z}$ has a predetermined rank typically provides worse results than imposing it directly in (2). Second, the inability to access the factorization of $\mathbf{Z}$ in (3) hinders the use of the "kernel trick". Third, (3) is a Semidefinite Program (SDP). Off-the-shelf SDP optimizers only scale to hundreds of variables, not amenable to the high dimensionality typically found in vision problems. While [8, 9, 24] ameliorate this issue, they still perform a SVD of $\mathbf{Z}$ in each iteration, making them unsuitable for handling dense, large scale datasets.

In this paper, we provide a best-of-all-worlds approach. Motivated by the theoretical results in [6], we show that many nuclear norm regularized problems of the form (3) can be optimized with a bilinear factorization of $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ by using the variational definition of the nuclear norm (see Fig. 1). While this reformulation is known in the literature, this paper is the first to propose a unification of traditional bilinear factorization and nuclear norm approaches under one formulation. This result allows us to analyze the conditions under which both approaches are equivalent and provide the best solution when they are not. Our analysis is divided in two situations: when the output rank is unconstrained and when the output rank is known *a priori*. For the first case, we propose a scalable and kernelizable optimization algorithm; for the second case, we propose a "rank continuation" strategy to avoid local optima. We show that our proposed strategies outperform state-of-the-art approaches in problems of Robust PCA, Photometric Stereo and Structure from Motion with outliers and missing data.

2. Previous Work

Low-rank matrix factorization is a long standing problem in computer vision. The seminal factorization method for Structure from Motion of Tomasi and Kanade [35] has been extended to encompass non-rigid and articulated cases, as well as photometric stereo [5] and multiple bodies [12]. Unfortunately, in the presence of missing data or weights, the factorization problem becomes NP-Hard [17]. Thus, many research works have focused on initialization strategies [21] or algorithms that are robust to initialization. Aguiar *et al*. [1] proposed an optimal algorithm in the absence of noise when the missing data follows a Young diagram. However, typical missing data scenarios in computer vision exhibit band patterns. Buchanan *et al*. [4] showed that alternated minimization algorithms are subject to flatlining and proposed a Newton method to jointly optimize $\mathbf{U}$ and $\mathbf{V}$. Okatani *et al*. [29] showed that a Wiberg marginalization strategy on $\mathbf{U}$ or $\mathbf{V}$ is very robust to initialization, but its high memory requirements make it impractical for medium-size datasets. These methods have also been extended to handle outliers [14, 15, 22]. De la Torre and Black [14] proposed a PCA with robust functions, and used Iterative Re-Weighted Least Squares (IRLS) to solve it. This approach can handle missing data by setting weights to zero in the IRLS algorithm; unfortunately, it is prone to local minima. Ke and Kanade [22] suggested replacing the LS loss with the L1 norm, minimized by alternated linear programming. Similarly to the LS case, Eriksson *et al*. [15] showed this approach is subject to flatlining and propose a Wiberg extension for L1. Wiberg methods have also been extended to arbitrary loss functions by Strelow [33], but exhibit the same scalability problems as its LS and L1 counterparts. Recently, Glashoff and Bronstein [18] proposed an Augmented Lagrange Multiplier (ALM) method for this problem, as a special case of [5] without constraints. This is subsumed by the generalized model proposed herein. Also, in Sec. 3 we provide a theoretical justification for this choice of algorithm for the factorization problem. The addition of problem specific constraints *e.g*., orthogonality of $\mathbf{U}$, has also been shown to help escape local minima in structure from motion [5, 41]. However, this is not generalizable to several other computer vision problems modeled as low-rank factorization problems [5, 31, 34, 37].

Alternatively to bilinear factorization approaches, Candes and Recht [10] stated that the rank function, under broad conditions of incoherence, can be minimized by its convex surrogate, the nuclear norm. This result has extended the use of low-rank priors to many applications where the rank is not known *a priori*, *e.g*., segmentation [11], background modeling [9] and tracking [39]. It has also been applied to penalize complexity in image classification and regression tasks [7, 20, 25]. Since the nuclear norm yields a SDP, several methods try to optimize it efficiently [8, 9, 16, 24] by exploiting a closed form solution of its proximal operator. However, they compute a SVD of $\mathbf{Z}$ in every iteration, hindering their applicability to large, dense datasets. Gradient methods on the Grassmann manifold have also been proposed to incrementally optimize (3) (*e.g*., [19]). However, these methods rely on a rank selection heuristic, which fails when the missing data pattern is not random.

Recently, the nuclear norm has also been proposed to tackle the factorization problem when the rank is known *a*

priori. Angst *et al*. [2] proposed a weighted version of the nuclear norm for structure from motion. Dai *et al*. [13] proposed an element-wise factorization for projective reconstruction by relaxing the rank 4 constraint to nuclear norm optimization. Zheng *et al*. [41] extended [15] by adding a nuclear norm regularizer to $\mathbf{V}$ and the orthogonality constraints in $\mathbf{U}$ found in [5] for structure from motion. Results provided in this paper shows that when the output rank is known *a priori*, nuclear norm solutions typically provide worse reconstruction results than those obtained with bilinear factorization models.

3. Soft and hard rank constraints

In this section, we bridge the gap between factorization and nuclear norm approaches. Consider the nuclear norm model in (3). For convex $f(\cdot)$, we can rewrite it as [32]

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{B},\mathbf{C}} \quad & f(\mathbf{X}-\mathbf{Z}) + \frac{\lambda}{2}\left(\mathrm{tr}(\mathbf{B})+\mathrm{tr}(\mathbf{C})\right) \\ \text{subject to} \quad & \mathbf{Q} = \begin{bmatrix} \mathbf{B} & \mathbf{Z} \\ \mathbf{Z}^\top & \mathbf{C} \end{bmatrix} \succeq 0. \end{aligned} \tag{4}$$

For any positive semidefinite matrix $\mathbf{Q}$, we can write $\mathbf{Q} = \mathbf{R}\mathbf{R}^\top$ for some $\mathbf{R}$. Thus, we can replace matrix $\mathbf{Q}$ in (4) by

$$\mathbf{Q} = \begin{bmatrix} \mathbf{B} & \mathbf{Z} \\ \mathbf{Z}^\top & \mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top & \mathbf{V}^\top \end{bmatrix}, \tag{5}$$

where $\mathbf{U} \in \mathbb{R}^{M\times r}$, $\mathbf{V} \in \mathbb{R}^{N\times r}$ and $r \leq \min(N, M)$ is an upper bound on $\mathrm{rank}(\mathbf{Z})$. Merging (5) into (4) yields

$$\min_{\mathbf{U},\mathbf{V}} \quad f(\mathbf{X}-\mathbf{U}\mathbf{V}^\top) + \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right), \tag{6}$$

where the SDP constraint was dropped because it is satisfied by construction. This reformulation seems counterintuitive, as we changed the convex problem in (3) into a non-convex one, which may be prone to local minima (*e.g*., in the case of missing data under the LS loss [17]). However, we show the existence of local minima in (6) depends only on the dimension r imposed on matrices $\mathbf{U}$ and $\mathbf{V}$. We extend the analysis of Burer and Monteiro [6] to prove that:

Theorem 1 *Let $f(\mathbf{X}-\mathbf{Z})$ be convex in $\mathbf{Z}$ and $\mathbf{Z}^*$ be an optimal solution of (3) with $\mathrm{rank}(\mathbf{Z}^*) = k^*$. Then, any solution $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ of (6) with $r \geq k^*$ is a solution of (3).*

Theorem 1 (which we prove in Appendix A) immediately allows us to draw one conclusion: The factorization and the nuclear norm models in (2) and (3) are special cases of (6). Fig. 2 illustrates this result in a synthetic case. We plot the output rank of $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ in (6) as a function of λ for a random 100×100 matrix $\mathbf{X}$ with all entries sampled i.i.d. from a Gaussian distribution $\mathcal{N}(0, 1)$, no missing data and

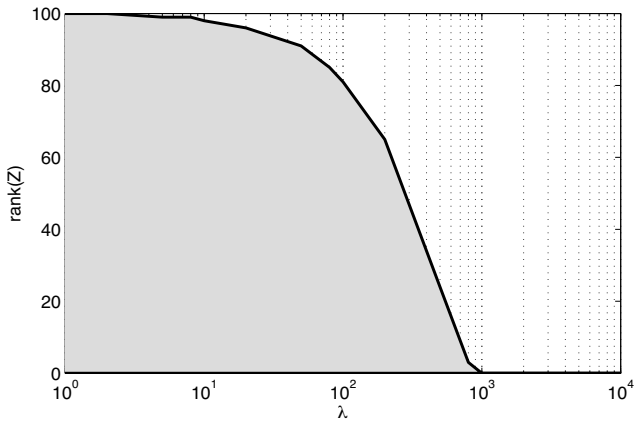


Figure 2. Region of equivalence between factorization (6) and nuclear norm approaches (3) for a 100×100 random matrix and LS loss. When factorization is initialized in the white area, it is equivalent to the result obtained with the nuclear norm (black line). When the rank is known *a priori*, better reconstruction results can be found in the grey area, by using factorization approaches.

LS loss: the factorization approach in (2) corresponds to the case where $\lambda = 0$ and r is fixed, whilst the nuclear norm in (3) outputs an arbitrary rank k^* as function of λ (the black line). According to Theorem 1 and the analysis in [6], for any $r \geq k^*$ (white area), (6) does not have additional local minima relative to (3). On the other hand, when $r < k^*$ (grey area), the conditions of Theorem 1 are no longer valid and thus optimizing (6) can be prone to local minima.

A special case of Theorem 1 has been used to recommend the use of nuclear norm approaches in the machine learning community by Mazumder *et al*. [27]. However, their analysis is restricted to the LS loss and the case where the rank is not known *a priori* (*i.e*., white area of Fig. 2). Our analysis instead extends to other convex loss functions and is based on the observation that many computer vision problems live in the grey area of Fig. 2. That is, their output rank k is predetermined by a domain-specific constraint (*e.g*., in Structure from Motion $k = 4$ [35]). Thus, we advocate the use of our unified model in (6) over the nuclear norm formulation in (3), based on two arguments:

When the output rank is unconstrained, we show in Sec. 4 that we can always choose $r \geq k^*$ such that (6) provides equivalent results to (3), while yielding scalable and kernelizable algorithmic solutions. In the case of the L1 loss, our algorithm for (6) has less computational complexity than exact proximal methods for (3).

When the output rank is known *a priori*, optimizing (6) is preferable to (2) and (3). As we will show in the experimental section, optimizing (6) is less prone to local minima than the unregularized problem (2). On the other hand, selecting λ in the nuclear norm model (3) such that the output rank k is the desired value typically leads to worse reconstructions than directly imposing $r = k$ in (6). Moreover, based on this analysis, we propose in Sec. 5 a "rank con-

tinuation" strategy, and empirically show it is able to attain global optimality in several scenarios.

For the remainder of the paper, we focus our attention in the LS loss $\|\mathbf{W}\odot(\mathbf{X}-\mathbf{Z})\|_F^2=\sum_{ij}(w_{ij}(x_{ij}-z_{ij}))^2$, and the L1 loss $\|\mathbf{W}\odot(\mathbf{X}-\mathbf{Z})\|_1=\sum_{ij}|w_{ij}(x_{ij}-z_{ij})|$, where $\mathbf{W}\in\mathbb{R}^{M\times N}$ is a weight matrix that can be used to denote missing data (*i.e.*, $w_{ij}=0$). We note, however, that this result applies to the Huber [2, 22] and hinge loss [25, 31].

4. Factorization with unconstrained rank

Nuclear norm models have extended the use of low-rank priors to many applications where $\mathbf{Z}$ is low rank but its exact value is not known *a priori* [9, 20, 40]. In this section, we propose an algorithm for solving (6) and show its complexity is lower than proximal methods [24] for optimizing the nuclear norm model in (3). One important factor to take into account when optimizing (6) is that when $\mathbf{U}$ or $\mathbf{V}$ are fixed, (6) becomes convex. However, it has been reported that pure alternation approaches are prone to flatlining [4, 15, 30]. For smooth losses such as the LS, this can be circumvented by performing gradient steps jointly in $\mathbf{U},\mathbf{V}$. Alternatively, we propose an Augmented Lagrange Multiplier (ALM) method for two reasons: 1) Theorem 1 and the analysis in [6] can be used to prove ALM's convergence to global optima of (3) when $r\geq k^*$, and 2) its applicability to the non-smooth L1 norm. Let us rewrite (6) as

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{U},\mathbf{V}}\quad & \|\mathbf{W}\odot(\mathbf{X}-\mathbf{Z})\|_1+\frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2\right)\\ \text{subject to}\quad & \mathbf{Z}=\mathbf{U}\mathbf{V}^\top,\end{aligned}\tag{7}$$

and its corresponding augmented lagrangian function as

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{U},\mathbf{V},\mathbf{Y},\rho}\quad & \|\mathbf{W}\odot(\mathbf{X}-\mathbf{Z})\|_1+\frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2\right)\\ & +\langle\mathbf{Y},\mathbf{Z}-\mathbf{U}\mathbf{V}^\top\rangle+\frac{\rho}{2}\|\mathbf{Z}-\mathbf{U}\mathbf{V}^\top\|_F^2,\end{aligned}\tag{8}$$

where $\mathbf{Y}$ are Lagrange multipliers and ρ is a penalty parameter to improve convergence [24]. This method exploits the fact that the solution for each subproblem in $\mathbf{U},\mathbf{V},\mathbf{Z}$ can be efficiently solved in closed form: For $\mathbf{U}$ and $\mathbf{V}$, the solution is obtained by equating the derivatives of (8) in $\mathbf{U}$ and $\mathbf{V}$ to $\mathbf{0}$. For known $\mathbf{U}$ and $\mathbf{V}$, $\mathbf{Z}$ can be updated by solving

$$\min_{\mathbf{Z}}\quad \|\mathbf{W}\odot(\mathbf{X}-\mathbf{Z})\|_1+\frac{\rho}{2}\|\mathbf{Z}-\left(\mathbf{U}\mathbf{V}^\top-\rho^{-1}\mathbf{Y}\right)\|_F^2,\tag{9}$$

which can be done in closed form by the element-wise shrinkage operator $\mathcal{S}_\mu(x)=\max(0,x-\mu)$, as

$$\begin{aligned}\mathbf{Z}=\mathbf{W}\odot\left(\mathbf{X}-\mathcal{S}_{\rho^{-1}}(\mathbf{X}-\mathbf{U}\mathbf{V}^\top+\rho^{-1}\mathbf{Y})\right)\\ +\overline{\mathbf{W}}\odot(\mathbf{U}\mathbf{V}^\top-\rho^{-1}\mathbf{Y}),\end{aligned}\tag{10}$$

for the L1 loss, or

$$\begin{aligned}\mathbf{Z}=\mathbf{W}\odot\left(\frac{1}{2+\rho}\left(2\mathbf{X}+\rho(\mathbf{U}\mathbf{V}^\top-\rho^{-1}\mathbf{Y})\right)\right)\\ +\overline{\mathbf{W}}\odot(\mathbf{U}\mathbf{V}^\top-\rho^{-1}\mathbf{Y}),\end{aligned}\tag{11}$$

for the LS loss. Here, $\overline{w}_{ij}=1,\forall_{ij}w_{ij}\neq 0$ and 0 otherwise. The resulting algorithm is summarized in Alg. 1.

Algorithm 1 ALM method for optimizing (6)

Input: $\mathbf{X},\mathbf{W}\in\mathbb{R}^{M\times N}$, params μ, λ, initialization of ρ
  while not converged **do**
    while not converged **do**
      Update $\mathbf{U}=(\rho\mathbf{Z}+\mathbf{Y})\,\mathbf{V}\left(\rho\mathbf{V}^\top\mathbf{V}+\lambda\mathbf{I}_r\right)^{-1}$
      Update $\mathbf{V}=(\rho\mathbf{Z}+\mathbf{Y})^\top\,\mathbf{U}\left(\rho\mathbf{U}^\top\mathbf{U}+\lambda\mathbf{I}_r\right)^{-1}$
      Update $\mathbf{Z}$ via (10) for L1 loss or (11) for LS loss
    end while
    $\mathbf{Y}=\mathbf{Y}+\rho(\mathbf{Z}-\mathbf{U}\mathbf{V}^\top)$
    $\rho=\min(\rho\mu,10^{20})$
  end while
Output: Complete Matrix $\mathbf{Z}=\mathbf{U}\mathbf{V}^\top$

Assuming without loss of generality that $\mathbf{X}\in\mathbb{R}^{M\times N}$ and $M>N$, we have that exact state-of-the-art methods for SVD (*e.g.*, Lanczos bidiagonalization algorithm with partial reorthogonalization) take a flop count of $O(MN^2+N^3)$. The most computational costly step in our ALM method are the matrix multiplications in the update of $\mathbf{U}$ and $\mathbf{V}$, which take $O(MNr+Nr^2)$ if done naively. Given that typically $k^*\leq r\ll min(M,N)$ and k^* can be efficiently estimated [23], Alg. 1 provides significant computational cost savings when compared to proximal methods which use SVDs [24].

Note on kernelization: An important implication of Theorem 1 is that in (6) we can solve (3) with access to $\mathbf{U},\mathbf{V}$. This makes kernel extensions trivial in *e.g.*, RLDA [20].

5. Factorization with rank known *a priori*

Many representations of shape, appearance or motion in computer vision problems yield models of a predetermined rank k [5, 12, 35]. In this case, we argue for the use of our model (6) instead of the nuclear norm approach in (3) and the unregularized model in (2). To understand why this is the case, let us consider the example of rank-k factorization of a matrix $\mathbf{X}$ under the LS loss with no missing data. For this case, both (2) and (3) have closed form solutions in terms of the SVD of $\mathbf{X}=\overline{\mathbf{U}}\overline{\Sigma}\overline{\mathbf{V}}^\top$, respectively, $\mathbf{Z}=\overline{\mathbf{U}}\overline{\Sigma}_{1:k}\overline{\mathbf{V}}^\top$ and $\mathbf{Z}=\overline{\mathbf{U}}\mathcal{S}_{\frac{\lambda}{2}}(\overline{\Sigma})\overline{\mathbf{V}}^\top$. In the case of noisy data, while the former yields the optimal rank-k reconstruction, we need to tune λ in the latter such that $\sigma_{k+1}=0$. If the λ required to satisfy this constraint is high, it may severely distort the non-zero singular values $\sigma_{1:k}$, resulting in poor reconstruction accuracy. On the other hand,

the analysis in Sec. 3 shows that rank restrictions typically lead to local minima when missing data are present. This problem is exacerbated when regularization is not used (*i.e.*, $\lambda = 0$): in addition to gauge freedom[2], it is clear that not all weight matrices $\mathbf{W}$ admit a unique solution [4]. As an extreme example, if $\mathbf{W} = \mathbf{0}$, any choice of $\mathbf{U}$ and $\mathbf{V}$ yields the same (zero) error. Thus, the unregularized factorization in (2) will be more prone to local minima than its regularized counterpart (6). These two arguments provide a general guideline for choosing λ: it should be selected as non-zero to ameliorate the local minima problem of (2), but small enough such that the first r singular values are not distorted.

Given that for any fixed λ we showed in Sec. 4 that (6) always has a region with no local minima, we propose the following "rank continuation" strategy: we initialize (6) with a rank $r \geq k^*$ matrix (*i.e.*, white region of Fig. 2), to guarantee its convergence to the global solution. Note that in the absence of an estimate for k^*, we can always use $r = \min(M, N)$. Then, we use this solution as initialization to a new problem (6) where the dimensions r of $\mathbf{U}, \mathbf{V}$ are decreased by one, until the desired rank is attained. This reduction can be done by using an SVD projection. This approach is summarized in Alg. 2.

Algorithm 2 Rank continuation

Input: $\mathbf{X}, \mathbf{W} \in \mathbb{R}^{M \times N}$, output rank k, parameter λ, an optional estimate of the output rank k^* of (3)
Initialize $\mathbf{U}, \mathbf{V}$ randomly, with $k^* \leq r \leq \min(M, N)$
Solve for $\mathbf{Z}$ in (6) with Alg. 1
for $r = \text{rank}(\mathbf{Z}) - 1, \ldots, k$ **do**
  SVD: $\mathbf{Z} = \overline{\mathbf{U}}\Sigma\overline{\mathbf{V}}^\top$
  Rank reduce: $\mathbf{U}_r = \overline{\mathbf{U}}\mathbf{\Sigma}^{\frac{1}{2}}_{1:r}$, $\mathbf{V}_r^\top = \mathbf{\Sigma}^{\frac{1}{2}}_{1:r}\overline{\mathbf{V}}^\top$
  Solve $\mathbf{Z}$ in (6) with initialization $\mathbf{U}_r, \mathbf{V}_r$ using Alg. 1
end for
Output: Complete Matrix $\mathbf{Z}$ with rank k

Rank continuation provides a *deterministic* optimization strategy that empirically is shown to find better local optima. We show in the experimental section that global minima of (6) are achieved with this strategy in several cases.

6. Experimental Results

In this section, we provide experimental validation of the conclusions drawn in Sec. 3, by applying our models in both synthetic and real datasets. We consider two computer vision scenarios: in Sec. 6.1, we explore the application of Robust PCA, a typical scenario when the desired output rank is unconstrained and compare our factorization (6) to nuclear norm approaches; in Sec. 6.2, we explore the applications of SfM, Non-rigid SfM and Photometric Stereo, typical scenarios where the output rank is known *a priori*. In this case, we compare our continuation approach to nuclear norm approaches and several factorization algorithms.

We use implementations provided in authors' websites for all baselines. For all experiments, we fix $\mu = 1.05$ initialize $\rho = 10^{-5}$ in Alg. 1. All experiments were run in a desktop with a 2.8 GHz Quad-core CPU and 6 GB RAM.

6.1. Factorization with unconstrained rank

In this section, we validated the lower computational complexity of the algorithm proposed in Sec. 4, when the output rank is unconstrained. We compared to state-of-the-art nuclear norm and Grassmann manifold methods: GRASTA [19], PRMF [36] and RPCA-IALM [24] in a synthetic and real data experiment for background modeling.

Synthetic data We mimicked the setup in [24] and generated low-rank matrices $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. The entries in $\mathbf{U} \in \mathbb{R}^{M \times r}, \mathbf{V} \in \mathbb{R}^{N \times r}$ and $M = N$ were sampled i.i.d. from a Gaussian distribution $\mathcal{N}(0, 1)$. Then, we corrupted 10% of the entries with large errors uniformly distributed in the range $[-50, 50]$. The error support was chosen uniformly at random. Like [24], we set $\lambda = \sqrt{N}$ and use the L1 loss. We varied the dimension N and rank r and measured the algorithm accuracies, defined as $\frac{\|\mathbf{Z}-\mathbf{X}\|_2}{\|\mathbf{X}\|_2}$, and the time they took to run. The results in Table 1 corroborate the analysis in Sec. 4: as N grows significantly larger than r, the smaller runtime complexity of our method allows for equally accurate reconstructions in a fraction of the time taken by RPCA-IALM. While PRMF and GRASTA are also able to outperform RPCA-IALM in time, these methods achieve less accurate reconstructions due to their alternated nature and sampling techniques, respectively.

Real data Next, we compared these methods on a real dataset for background modeling. Here, the goal is to obtain the background model of a slowly moving video sequence. Since the background is common across many frames, the matrix concatenating all frames is a low rank matrix plus a sparse error matrix modeling the dynamic foreground.

We followed the setup of [36] and used the Hall sequence[3]. This dataset consists of 200 frames of video with a resolution of 144×176, and we set the scope of the virtual camera to have the same height, but half the width. We simulated a camera panning by shifting 20 pixels from left to right in frame 100 to simulate a dynamic background. Additionally, we randomly dropped 70% of the pixels. We proceeded as in the previous synthetic experiment. Fig. 3 shows a visual comparison of the reconstruction of several methods. Results corroborate the experiment in Tab. 1 and show that the lower accuracies of GRASTA and PRMF yield noisier reconstructions than our method.

[2] for each solution $\mathbf{U}\mathbf{V}^\top$, any solution $(\mathbf{U}\mathbf{R})(\mathbf{R}^{-1}\mathbf{V}^\top)$ where $\mathbf{R} \in \mathbb{R}^{r \times r}$ is an invertible matrix will provide an equal cost.

[3] http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html

Table 1. Performance comparison of state-of-the-art methods for Robust PCA. Time is in seconds. Error has a factor of 10^{-8}.

Matrix		RPCA-IALM [24]		GRASTA [19]		PRMF [36]		Ours	
N	r	Error	Time	Error	Time	Error	Time	Error	Time
100	3	1.4872	0.3389	226.46	1.7656	3338.7	0.4704	**0.5286**	**0.1734**
200	5	1.5599	2.3575	241.99	2.7282	2687.5	1.0382	**0.7182**	**0.5739**
500	10	3.2595	10.501	263.55	9.5399	1692.4	6.2480	**0.1273**	**3.2373**
1000	15	0.3829	44.111	286.17	23.535	1145.8	30.441	**0.0701**	**14.339**
2000	20	0.6212	196.89	329.11	83.010	808.20	126.95	**0.0308**	**60.658**
5000	25	0.2953	1840.0	379.94	**507.57**	504.08	1307.4	**0.0589**	556.21

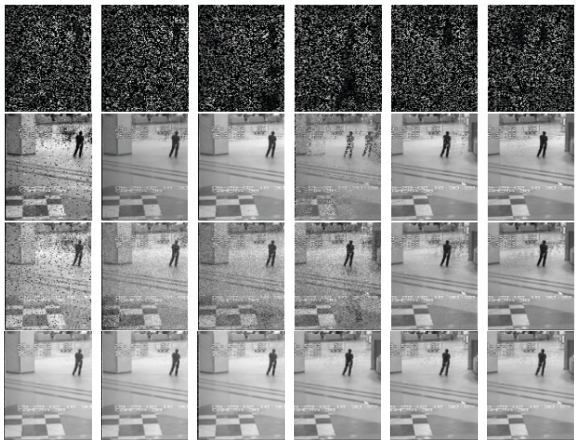
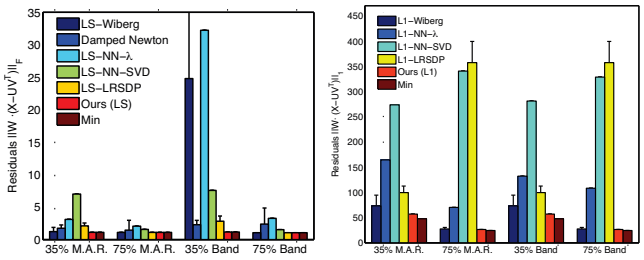


Figure 3. Results for background modeling with virtual pan. The first row shows the known entries used for training in frames 40, 70, 100, 130, 170, 200. The remaining rows show the results obtained by PRMF, GRASTA and our method, respectively.

6.2. Factorization with known rank

In this section, we empirically validated the "rank continuation" strategy proposed in Sec. 5, in several synthetic and real data problems where the output rank is known *a priori*. We compared our method to state-of-the-art factorization approaches: the damped Newton in [4], the LRSDP formulations in [28] and the LS/L1 Wiberg methods in [15, 29]. Following results reported in the detailed comparisons of [4, 15, 29, 30, 33, 41], we dismissed alternated methods due to their flatlining tendency. To allow direct comparison with published results [4, 28, 29, 41], all methods solved either (2) or (6) without additional problem specific constraints and we fixed $\lambda = 10^{-3}$. For control, we also compared to two nuclear norm baselines: NN-SVD, obtained by solving (3) with the same λ used for other models and projecting to the desired rank with an SVD; NN-λ, obtained by tuning λ in (3) so the desired rank is obtained.

Synthetic data We assessed the convergence performance of our continuation strategy using synthetic data. We performed synthetic comparisons for the two loss choices described in Sec. 3: LS loss and L1 loss. For the LS loss, we generated rank-3 matrices $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. The entries in $\mathbf{U} \in \mathbb{R}^{20\times 3}$, $\mathbf{V} \in \mathbb{R}^{25\times 3}$ were sampled i.i.d. from a Gaussian distribution $\mathcal{N}(0,1)$ and Gaussian noise $\mathcal{N}(0,0.1)$ was added to every entry of $\mathbf{X}$. For the L1 loss, we proceeded as described for the LS case but additionally corrupted 10% of the entries chosen uniformly at random with outliers uniformly distributed in the range $[-2.5, 2.5]$. We purposely kept the synthetic experiments small, due to the significant memory requirements of the Wiberg algorithms. We varied the percentage of known entries and measured the residual over all *observed* entries, according to the optimized loss function. We chose this measure as it allows for direct comparison between unregularized and regularized models. We ran damped Newton, LRSDP and Wiberg methods 100 times for each test with random initializations.



(a) LS loss (b) L1 loss

Figure 4. Comparison of convergence to empirical global minima (Min) for the LS and L1 losses in synthetic data. The minima are found as the minimum of all 100 runs of all methods for each test.

Fig. 4 shows the results for the LS and L1 loss cases. We show two representatives cases for the percentage of known entries (75% and 35%, the breakdown point for L2-Wiberg methods), both for missing data patterns at random (M.A.R.) and with a pattern typical of SfM matrices (Band), generated as in [29]. The theoretical minimum number of entries to reconstruct the matrix is the same as the number of parameters minus factorization ambiguity $Mr+(N-r)(r+1)$, which for this case is 29.6% [29]. We verified the behavior of all methods when more than 40% of the entries are known is similar to the result shown for 75%.

For the LS case, results in Fig. 4(a) show that our deterministic continuation approach always reaches the empirical optima (found as the minimum of all runs of all methods), regardless of the number of known entries or pattern of missing data. Note the minimum error is not zero, due to the variance of the noise. As reported previously [29, 30, 41],

Table 2. Real datasets for problems with known output rank-k.

Dataset	Size	Output rank k	Known entries
Dino	319×72	4	28%
Giraffe	240×167	6	70%
Face	2944×20	4	58%
Sculpture	26260×46	3	41%

we observe that L2-Wiberg is insensitive to initialization for a wide range of missing data entries. However, we note that its breakdown point is not at the theoretical minimum, due to the lack of regularization. The LRSDP method for optimizing (6) outperforms the Wiberg method, suggesting that similar convergence properties of the Wiberg can be obtained without its use of memory. The baseline NN-SVD performed poorly, showing that the estimation of the nuclear norm fits information in its additional degrees of freedom instead of representing it with the true rank. For the L1 loss case, results in Fig. 4(b) show that our continuation strategy no longer attains the empirical optima. We note that this is not surprising since **the problem of factorization with missing data is NP-Hard**. However, its deterministic result is very close to the optima. Our continuation method regained empirical optimality when only 2% of outliers were present in the data, suggesting a dependency on the noise for the L1 case. In this case, our performance is comparable to what is obtained with the L1-Wiberg algorithm [15] on average. Thus, continuation is a viable alternative to the memory expensive Wiberg method.

Real data Next, we assessed the results of our continuation approach in real data sequences. We used four popular sequences[4]: a) Dino, for affine SfM; b) Giraffe, for non-rigid SfM, and c) Face and d) Sculpture, both photometric stereo sequences. Their details are summarized in Table 2. The dimension of these datasets make the usage of the L1-Wiberg of [15] prohibitive in our modest workstation, due to its memory requirements. For the Sculpture dataset, we treated as missing all pixels with intensity greater than 235 or lower than 20 (*e.g.*, in Fig. 5(b), the yellow and purple+black masks, resp.). All other datasets provide $\mathbf{W}$.

Table 3 shows a comparison of average error over all *observed* entries for the continuation proposed in Alg. 2 and several methods, according to the loss functions L1/LS. "Best" denotes the best known result in the literature. As explained in Sec. 5, we observe that nuclear norm regularized approaches NN-SVD and NN-λ result in bad approximations when a rank restriction is imposed. Similar to the results in the synthetic tests, our method always attained or outperformed the state-of-the-art result for the LS loss. The convergence studies in [4, 29] performed optimization on the first three datasets several times with random initializations, so their reported results are suspected by the community to be the global optima for these problems. Our method consistently attains these results in a deterministic fashion,

[4]http://www.robots.ox.ac.uk/~abm/

(a) Input (b) Mask (c) NN-λ (d) RL1NN (e) NNSVD (f) Ours

Figure 5. Results for frame 17 of the sculpture sequence. While (c) (d) smooth out the image and (e) fails to reconstruct it, our continuation approach (f) is able to obtain reconstructions preserve finer details, such as the imperfections on the cheek or chin.

Table 3. Comparison of LS/L1 average error over all *observed* entries for structure from motion and photometric stereo datasets.

Dataset	$f(\cdot)$	Best	NN-λ	NN-SVD	Ours
Dino	LS	**1.0847** [4]	6.1699	35.8612	**1.0847**
	L1	0.4283 [41]	7.6671	80.0544	**0.2570**
Giraffe	LS	**0.3228** [4]	0.4370	0.6519	**0.3228**
	L1	–	1.8974	11.0196	**0.2266**
Face	LS	**0.0223** [4]	0.0301	0.0301	**0.0223**
	L1	–	0.0287	0.6359	**0.0113**
Sculpt	LS	24.6155 [5]	44.5859	31.7713	**22.8686**
	L1	17.753 [41]	21.828 [38]	33.7546	**12.6697**

as opposed to state-of-the-art methods which get stuck in local minima several times. As a control experiment, we also ran our continuation strategy for the unregularized case ($\lambda = 0$) on the Dino sequence with LS loss, which resulted in a RMSE of 1.2407. We attribute this to the fact that this case is more prone to local minima, as mentioned in Sec. 5.

For the L1 loss, continuation outperforms the state-of-the art in all datasets. It might be argued that problem specific constraints are required to obtain clean reconstructions, but we reiterate the importance of escaping local minima. While there are certainly degenerate scenarios which can only be solved with such constraints [26], Alg. 1 (and consequently, Alg. 2) can be trivially extended to handle such cases. For example, the projection step on $\mathbf{U}$ for SfM in [5] can be added to Alg. 1 or the problem can be reformulated as a different SDP [28] with a rank constraint, which can be tackled by our continuation strategy in Alg. 2.

7. Conclusion

We developed a unified approach to matrix factorization and nuclear norm regularization, that inherits the benefits of both approaches. Based on this analysis, we proposed a deterministic "rank continuation" strategy that outperforms state-of-the-art approaches in several computer vision applications with outliers and missing data. Future work in factorization algorithms should optimize the unified model in (6), since it subsumes the traditional factorization and the nuclear norm regularized approaches. It should also focus into the theoretical understanding of the continuation model proposed in Sec. 5, as both synthetic and real data experimental results provide a strong indication of it achieving the

global minima when using the LS loss.

Acknowledgements: Support for this research was provided by the Portuguese Foundation for Science and Technology through the Carnegie Mellon Portugal program under the project FCT/CMU/P11. Partially funded by FCT projects Printart PTDC/EEA-CRO/098822/2008 and PEst-OE/EEI/LA0009/2013 and the Poeticon++ project from the European FP7 program (grant agreement no. 288382). Fernando De la Torre is partially supported by Grant CPS-0931999 and NSF IIS-1116583. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

A. Proof of Theorem 1

To show this, we first note that (6) agrees with the following alternative formulation of the nuclear norm [32],

$$\|\mathbf{Z}\|_* \quad = \min_{\mathbf{Z}=\mathbf{U}\mathbf{V}^\top} \frac{1}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right). \qquad (12)$$

and that Mazumder *et al*. [27] showed the following result:

Lemma 2 *For any* $\mathbf{Z} \in \mathbb{R}^{M\times N}$*, the following holds: If* $\text{rank}(\mathbf{Z}) = k \leq \min(M, N)$*, then the minimum of* (12) *is attained at a factor decomposition* $\mathbf{Z} = \mathbf{U}_{M\times k}\mathbf{V}_{N\times k}^\top$.

This result allows us to prove the desired equivalence:

Proof Applying Lemma 2, we can reduce (6) to

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} && f(\mathbf{X} - \mathbf{U}\mathbf{V}^\top) + \lambda\|\mathbf{U}\mathbf{V}^\top\|_* \\ =&\min_{\mathbf{Z},\text{rank}(\mathbf{Z})=k} && f(\mathbf{X} - \mathbf{Z}) + \lambda\|\mathbf{Z}\|_* \\ =&\min_{\mathbf{Z}} && f(\mathbf{X} - \mathbf{Z}) + \lambda\|\mathbf{Z}\|_*. \end{aligned} \qquad (13)$$

References

[1] P. Aguiar, J. Xavier, and M. Stosic. Spectrally optimal factorization of incomplete matrices. In *CVPR*, 2008.
[2] R. Angst, C. Zach, and M. Pollefeys. The generalized trace-norm and its application to structure-from-motion problems. In *ICCV*, 2011.
[3] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.
[4] A. M. Buchanan and A. W. Fitzgibbon. Damped newton algorithms for matrix factorization with missing data. In *CVPR*, 2005.
[5] A. D. Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear Modelling via Augmented Lagrange Multipliers (BALM). *PAMI*, 2012.
[6] S. Burer and R. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Math. Prog.*, 103(3):427–444, 2005.
[7] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino. Matrix completion for multi-label image classification. In *NIPS*, 2011.
[8] J.-F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 20(4):1956–1982, 2008.
[9] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011.
[10] E. Candès and B. Recht. Exact low-rank matrix completion via convex optimization. In *Allerton Conference*, 2008.
[11] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan. Multi-task low-rank affinity pursuit for image segmentation. In *ICCV*, 2011.
[12] J. P. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *IJCV*, 29(3):159–179, 1998.
[13] Y. Dai, H. Li, and M. He. Element-wise factorization for n-view projective reconstruction. In *ECCV*, 2010.
[14] F. De La Torre and M. J. Black. A Framework for Robust Subspace Learning. *IJCV*, 54(1-3):117–142, 2003.
[15] A. Eriksson and A. Hengel. Efficient computation of robust low-rank matrix approximations in the presence of missing data using the L1 norm. In *CVPR*, 2010.
[16] P. Favaro, R. Vidal, and A. Ravichandran. A closed form solution to robust subspace estimation and clustering. In *CVPR*, 2011.
[17] N. Gillis and F. Glineur. Low-Rank Matrix Approximation with Weights or Missing Data Is NP-Hard. *SIAM Journal on Matrix Analysis and Applications*, 32(4), 2011.
[18] K. Glashoff and M. M. Bronstein. Structure from motion using augmented lagrangian robust factorization. In *3DIMPVT*, 2012.
[19] J. He, L. Balzano, and A. Szlam. Incremental gradient on the grassmannian for online foreground and background separation in subsampled video. In *CVPR*, 2012.
[20] D. Huang, R. Cabral, and F. De la Torre. Robust regression. In *ECCV*, 2012.
[21] D. W. Jacobs. Linear fitting with missing data for structure-from-motion. *CVIU*, 82:206–2012, 1997.
[22] Q. Ke and T. Kanade. Robust l_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *CVPR*, 2005.
[23] R. H. Keshavan and S. Oh. A gradient descent algorithm on the grassman manifold for matrix completion. *arXiv*, 0910.5260, 2009.
[24] Z. Lin, M. Chen, and Y. Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *Mathematical Programming*, 2010.
[25] N. Loeff and A. Farhadi. Scene discovery by matrix factorization. In *ECCV*, 2008.
[26] M. Marques and J. Costeira. Estimating 3d shape from degenerate sequences with missing data. *CVIU*, 113(2):261–272, 2009.
[27] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization for learning large incomplete matrices. *JMLR*, 99:2287–2322, 2010.
[28] K. Mitra, S. Sheorey, and R. Chellappa. Large-scale matrix factorization with missing data under additional constraints. In *NIPS*, 2010.
[29] T. Okatani and K. Deguchi. On the Wiberg algorithm for factorization with missing components. *IJCV*, 72(3):329–337, 2007.
[30] T. Okatani, T. Yoshida, and K. Deguchi. Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms. In *ICCV*, 2011.
[31] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Bilinear classifiers for visual recognition. In *NIPS*, 2009.
[32] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, Aug. 2010.
[33] D. Strelow. General and nested Wiberg minimization. In *CVPR*, 2012.
[34] J. B. Tenenbaum and W. T. Freeman. Separating Style and Content with Bilinear Models. *Neural Computation*, 1283:1247–1283, 2000.
[35] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9:137–154, 1992.
[36] N. Wang, T. Yao, J. Wang, and D. Yeung. A probabilistic approach to robust matrix factorization. In *ECCV*, 2012.
[37] J. Warrell, P. Torr, and S. Prince. StyP-Boost : A Bilinear Boosting Algorithm for Style-Parameterized Classifiers. In *BMVC*, 2010.
[38] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *ACCV*, 2010.
[39] F. Xiong, O. I. Camps, and M. Sznaier. Dynamic context for tracking behind occlusions. In *ECCV*, 2012.
[40] Z. Zhang, Y. Matsushita, and Y. Ma. Camera calibration with lens distortion from low-rank textures. In *CVPR*, 2011.
[41] Y. Zheng, G. Liu, S. Sugimoto, S. Yan, and M. Okutomi. Practical low-rank matrix approximation under robust l1-norm. In *CVPR*, 2012.