This ICCV workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Efficient and Robust Inverse Lighting of a Single Face Image using Compressive Sensing

Miguel Heredia Conde[†], Davoud Shahlaei[#], Volker Blanz[#] and Otmar Loffeld[†]
Center for Sensor Systems[†] (ZESS) and Institute for Vision and Graphics[#], University of Siegen
57076 Siegen, Germany
heredia@zess.uni-siegen.de

Abstract

In this paper, we show that the recent theory of Compressive Sensing (CS) can successfully be applied to solve a model-based inverse lighting problem for single face images, even in harsh lighting with multiple light sources, including cast shadows and specularities. It has been shown that an illumination cone can be used to perform realistic inverse lighting. In this work, the cone images are synthetically generated using directional lights and a realistic reflectance of faces. Thereby, the face model is achieved by fitting a 3D Morphable Model to the input image. We apply CS to find the sparsest illumination setup from few random measurements of the RGB input and the cone images. The proposed method significantly reduces the dimensionality through stochastic sampling and a greedy algorithm for the sparse support estimation, yielding low runtimes. The greedy search is designed to handle non-negativity of the light sources and joint-support selection. We show that the proposed method reaches a quality of illumination estimation equal to previous work, while dramatically reducing the number of active light sources. Thorough experimental evaluation shows that stable recovery is achievable for compression rates up to 99%*. The method exhibits outstanding robustness to additive noise in the input image.*

1. Introduction

This work brings the areas of inverse lighting of faces and Compressive Sensing (CS) together. We show that CS can profit from the low number of active illumination sources needed to recreate the illumination conditions in the input image to efficiently solve the problem.

1.1. Inverse Lighting

Single image inverse lighting aims for estimation of lighting parameters for a given rendering machine to render an image as close as possible to the input image. For rendering, not only lighting but also parameters such as 3D shape, albedo and reflectance of the surfaces, camera properties and etc. are necessary. We specifically refer to the realistic inverse lighting process proposed by Shahlaei and Blanz [27]. The focus is on human faces, as one of the most attractive real objects for computer vision and graphics. This is the pipeline:

1. Estimate face model by fitting a 3D Morphabel Model (3DMM) [4] to the input image.
2. Configure the renderer with the estimated 3D geometry, together with a generic realistic reflectance and texture for faces based on measurements of [29].
3. Using the renderer and a fixed number of light sources, render images of the face with each single light.
4. Find an optimum coefficient vector for combining the rendered images, so that the weighted summation becomes as similar as possible to the input image. (See Eq. 1.)

They call the set of rendered images a *synthetic illumination cone*, comparing it to an *illumination cone* which is extensively defined in [3]. One successful approach with a real illumination cone is to build an orthogonal basis out of the cone images with spherical harmonics [25]. Conversely, we directly take the rendered cone images as the basis. The use of a realistic representation of lighting with non-negative lighting parameters leads to realistic reconstruction of illumination effects such as specular and Fresnel highlights and cast shadows, which are perceived strongly in real images. When combined with a realistic reflectance function and a 3D model, even complex unknown lightings can be estimated from a single face image. On the downside, we have to deal with the linear dependencies among basis elements. Moreover, using a synthetic cone for inverse lighting of a captured image introduces even more challenges, specially when the only available data is that single image. According to [27], these challenges rise from all the unknowns, such as

shape and texture of the face, reflectance of the skin, camera properties and the color balance of the image. Based on superposition principle, light is additive. In illumination cone terms, it means that any lighting situation can be reconstructed as a weighted sum of the cone images. It has been shown that an environment map, representing the incident light for an object, can be summarized in a few light sources for rendering of that object [12]. Therefore, it is possible to formulate the superposition as below:

$$\vec{I} = \mathbf{\Psi}\vec{X} \tag{1}$$

where $\vec{I}$ the vector of face pixels in the input image and $\vec{X}$ is the coefficient vector for the linear combination of cone images in matrix $\mathbf{\Psi}$.

To find the lighting parameters means to correctly find a non-negative vector $\vec{X}$ for the Eq. 1. This paper shows how to successfully apply Compresive Sensing to solve this real life problem.

1.2. Compressive Sensing

Few research areas have experienced a development as fast as that of Compressive Sensing (CS). The success on applying CS to very different areas of science testifies the evolution from a mathematical theory to an useful tool that can be used in cases where the high dimensionality of the data is an issue, typically in data acquisition and transfer. CS proposes to apply the compression at sensing and not after, that is, to gather few measurements, less than those required for the desired signal resolution, but still containing all signal information. Fundamental works in CS [5, 15] have shown that the high-resolution signal can be recovered from the measurements if it is sparse (or, at least, compressible) in a certain dictionary and certain conditions on the sensing process and its relation to the dictionary are fulfilled. The number of measurements required to recover a signal is directly related to its sparsity in the dictionary and not to the maximum frequency contained in it. Consequently, CS theory implicitly suggests that highly-sparse signals can be recovered from much less measurements than those suggested by the Nyquist rate.

At first glance, the topics of realistic inverse lighting and CS might seem unrelated, but a closer look reveals that the previous can take enormous profit of the novel sensing paradigm proposed by the latter. It is clear that the amount of information contained in an image does not necessarily grow with its resolution. Specifically, in our case, the amount of *useful* information does not grow at all after some low resolution is reached, since illumination effects are of low spatial frequency. Higher resolutions are not advantageous in information terms. Additionally, large images might lead to intractable optimization problems when operating directly in pixel domain. CS natively accounts for such issues, operating in a compressed domain, which is assured to preserve the information contained in the original image if it admits a sparse representation in an appropriate dictionary. This bring us to the next strong point of this work, which is the signal sparsity. Most natural signals are not exactly sparse in general dictionaries such as orthogonal basis, but compressible, e.g., periodic signals in Fourier domain or natural images in wavelet domain. This means that part of the signal power is to be distributed among many or all dictionary atoms, since the signal cannot be exactly represented as a sparse linear combination of them. We show that the inverse lighting problem falls into the category of purely sparse signal recovery, since very few active light sources are typically needed to approximate the illumination in the input face image.

2. Related Work

2.1. Inverse Lighting

Inverse rendering methods can be divided in two groups based on the type of the input. On the one hand, a group of inverse rendering approaches acquire the appearance of the given face with a configuration of multiple cameras, 3D scanners and geodesic light domes. Some of these works, including [3, 13, 18, 29] measure the facial appearance under different lighting situations in a lab, in several 2D images or 3D scans, and use the results to estimate the illumination. Some other data driven approaches, such as [17], infer information about the lighting of the scene with the help of a reference object. Providing visually impressive results, this group have demonstrated the challenges of appearance acquisition and inverse lighting in a controlled environment. They also have shown that it is challenging to reconstruct perfect results when it comes to the human skin.

On the other hand, there are single image methods which make assumptions on unknown parameters or estimate them to perform inverse rendering for faces [4, 26, 1, 19, 27]. Among these methods, using a 3D morphable model, to estimate shape and texture of the face, has gained attention. While some of them [4, 19, 26] fit the whole rendering parameters (i.e., shape, texture, lighting, color balance, *etc.*) in a unified process, some others [1, 27] perform an extra illumination estimation step to refine the results.

The single image inverse rendering methods demand no special hardware and are free from the presence of the person in the lab. This adds to the flexibility and the amount of imaginable use cases. As a tradeoff, they usually lack visual quality compared to the first group. While many aspects of the appearance acquisition methods in the first group have been researched, the second group is still dealing with a hard kind of open problem.

Using illumination cone is common in both groups. All illumination cone methods, but [27], use non-physical representation of lights, such as Spherical Harmonics, to build an orthogonal basis from a captured illumination cone. A

mathematically manipulated basis has clear mathematical benefits, yet, fails in reconstruction of some complexities of natural lighting effects. Usually specular highlights, and cast shadows are the most important effects that are being compromised. However, a physically plausible basis lacks orthogonality.This demands mathematical approaches which are specifically configured to deal with the ambiguity of the model. On the bright side, a this approach assures to take all the physical effects of light into account and shows more stability against complexity of the lighting and more potential of improvement, whenever other parameters of rendering are estimated more accurately.

2.2. Compressive Sensing

The basic theoretical background underlying this work is Compressive Sensing (CS). Introductory works on CS are [2, 7]. While the basis of modern CS can be found in, e.g., [8, 15], a good compendium of recent developments and applications is given in [16]. The core idea of CS is that a signal which admits a sparse or compressible representation can be recovered from a reduced set of non-adaptive measurements. In the case of an exactly sparse signal, an accuracy similar to that achievable with knowledge of the sparse support of the signal is guaranteed if a minimum number of measurements is performed. This already introduces the two main fields of CS: how the measurements are to be performed and how to recover the sparse vector of coefficients from the measurements, given the measurement strategy and the sparsity dictionary.

Regarding to the sensing process, Fourier matrices [5], Gaussian random matrices [9] and Bernoulli binary matrices have been shown to allow for sparse recovery. Recovery rates and error bounds for these matrices are given in [6]. Binary sensing matrices allow for fast computation and high efficiency in an eventual hardware implementation.

In [5, 8, 15] it was stated that $m = O(k \log(n))$ measurements are necessary for sparse signal recovery, that is, the minimum number of measurements required for recovery depends linearly on the sparsity, k and only logarithmically on the signal size, n. This is of capital importance in this work, since it allows our method to perform a dramatic dimansionality reduction, backed by the very low values of k (number of active light sources) and to deal with large images directly, without prior downsampling.

Finding the sparsest solution to the system of linear equations (SLE) in Eq. 5 is equivalent to minimizing the l_0 norm of the solution, subject to the constraints imposed by the SLE. It has been shown that finding a solution to such a problem is, in general, NP-hard [22]. In the literature, there are two different approaches to overcome this difficulty: to substitute the l_0 minimization by an l_1 minimization or to design a *greedy* algorithm to find the sparse support of the signal and estimate the corresponding non-zero coefficients of the sparse solution.

In [10] it is shown that such linearly-constrained l_1 minimization can be efficiently solved as a linear program, even when dealing with overcomplete dictionaries. While convex optimization methods for l_1 minimization are a standard way of signal recovery in CS, greedy algorithms deserve special mention, since they allow for lower complexity and greater flexibility than convex optimization, leading to very reduced runtimes in cases of high sparsity.

The general structure of a greedy algorithm consists of two steps: support element selection and coefficients update. A basic pursuit algorithm is the Matching Pursuit (MP), presented in [20], where it was already shown that such greedy pursuits can operate in overcomplete dictionaries and deal natively with noise, which is filtered out not because of its lower level, but because of its incoherence with the dictionary atoms. MP selects a single column of $\boldsymbol{A}$ at each iteration and only the associated coefficient is updated. The Orthogonal Matching Pursuit (OMP) [24], also known as the Orthogonal Greedy Algorithm [14], adds a new element to a support set at each iteration and X is updated by projecting Y orthogonally onto the columns of $\boldsymbol{A}$ indexed by that support set. OMP selects at each iteration the dictionary atom that is most correlated with the residual, but this does not imply selecting the atom that provides the largest error reduction after orthogonalization. The Order Recursive Matching Pursuit (ORMP) [23] solves this issue with the introduction of an orthogonal projection operator, which is updated for the new support set in each iteration. This way, ORMP selects, at each iteration, the support element whose normalized projection orthogonal to the subspace spanned by the columns of A indexed by the current support set is maximally correlated to the current residual.

2.3. Compressive Sensing and Inverse Lighting

In contrast to the mainstream research on inverse lighting, we propose the use of Compressive Sensing (CS) to deal with basis redundancy and provide a solution which minimizes the number of active light sources. Although this approach is thematically close to some previous work [21, 31], our method deals with much more complicated lighting conditions; it is invariant to pose and camera, considers and replicates realistic reflectance effects and works natively with color images. Such degree of freedom in the proposed method contributes to the complexity of the problem and demands special care in the design of the solution.

Additionally, the classification of sparse estimation methods in [21, 31] as CS is doubtful, since the required stochastic sampling is missing. In [21] spherical harmonics are used as measurement method, providing a large dimensionality reduction, but at the cost of bounding the range of illumination effects that can be recovered. Using random sensing matrices does not compromise the information and

meets the incoherence between sensing and representation demanded by CS. In [31] no dimensionality reduction takes place and their Sparse Representation-based Classification (SRC) is therefore based on l_1 minimization with constrains in the high-dimensional image space.

A crucial observation was made in [30], where random projections and downsampled images where used as image features and it was stated that, if a sparse representation is possible, the choice of features, i.e., the sensing method, is irrelevant and what matters is its number, i.e., the number of *compressed* measurements. Although the goal of [30] is different to ours, it is related as it actually performs a *compressed* sensing prior to sparse recovery. Similarly, we also make use of random projection, with and without prior downsampling, as sensing method.

3. Optimization with Non-Negative Least Squares

In this section we briefly describe the optimization method from [27], which is taken as a reference for comparison to the proposed method in Section 5. A Pseudo-Newton-Raphson iterative algorithm successfully minimizes a sum of least squares cost function (Eq. 2) to find the optimum coefficient vector per channel. Let $\Gamma = \{R, G, B\}$ be the set of image channels, consider the matrix $\boldsymbol{X} = [\vec{X^\gamma}]_{\gamma\in\Gamma}$ constructed stacking coefficient vectors by columns. Eq. 2 comes from Eq. 1, considering images in RGB color space and the color balance differences between the rendered images $\boldsymbol{\Psi_i} = [\vec{\psi_i^\gamma}]_{\gamma\in\Gamma}$ and the input image $\boldsymbol{I} = [\vec{I^\gamma}]_{\gamma\in\Gamma}$ with a transformation matrix $\boldsymbol{T}$ and an offset vector $\vec{o} = [o^\gamma]_{\gamma\in\Gamma}$.

$$\overrightarrow{E(\boldsymbol{X})} = \frac{1}{2}\sum_{p\in face}\left(\vec{o} + \boldsymbol{T}(\overrightarrow{\sum_{i=1}^{n_{atoms}}\left(\vec{x_i}\,.\,\vec{\psi_i}(p)\right)}) - \vec{I(p)}\right)^2 \quad (2)$$

where, $\vec{I(p)}$ represents the R, G and B values of pixel p from the input image, $\vec{o}$ and $\boldsymbol{T}$ (Eq. 3) parameters for color correction of the renderd images with respect to the input image, $\vec{x_i}$ the i-th RGB coefficient and $\vec{\psi_i}(p)$ the RGB values of pixel p of the i-th cone image, n_{atoms} the number of cone images, and therefore, the $\overrightarrow{(\sum_{i=1}^{n_{atoms}}\left(\vec{x_i}\,.\,\vec{\psi_i}(p)\right))}$ is the weighted linear combination of the cone images at pixel p. The color balance parameters are based on the model described in [27], which transforms the neutral colors of the model to the color contrast of the input image. To control the overall brightness, the offset vector $\vec{o}$ is added to the pixels of a rendered image after transformation $\boldsymbol{T}$ has been applied to it:

$$\boldsymbol{T} = \mathrm{diag}(\vec{g})\begin{pmatrix}(0.7\xi+0.3) & (0.6-0.6\xi) & (0.1-0.1\xi)\\ (0.3-0.3\xi) & (0.4\xi+0.6) & (0.1-0.1\xi)\\ (0.3-0.3\xi) & (0.6-0.6\xi) & (0.9\xi+0.1)\end{pmatrix} \quad (3)$$

where $\vec{g} = [g^\gamma]_{\gamma\in\Gamma}$ the estimated gain for each channel and ξ is the estimated color contrast. This transformation not only controls the gain of each color channel separately, but also allows a ξ-weighted interpolation between the gray value and the saturated value of each pixel. The offsets, gains and the value of ξ are estimated during the fitting of the 3DMM to the input image [4]. Please note that the $\vec{\Psi}_i$ images are rendered with $\xi = 1$, $\vec{o} = \vec{0}$ and $\vec{g} = \vec{1}$, so that they have the neutral saturation of the rendering machine. Modeling the color balance is necessary for the estimation of illumination and reconstruction of scenes when dealing with differently saturated input images images.

The cost function in Eq. 2 is regularized with a prior on the value of $\vec{X}$ for all the 3 color channels in Eq. 4.

$$r(\boldsymbol{X}) = \eta\left(\textstyle\sum_{\gamma\in\Gamma}(\sum_{i=1}^{n_{atoms}}\frac{(x_i^\gamma)^2}{\sigma_i^2})\right) \quad (4)$$

This regularization term is tuned with η and σ_i, $\forall i \mid i \in [1, n_{atoms}]$. It penalizes all negative and positive values and forces each x_i^γ to become zero, depending on their influence on the value of the cost function. This l_2 regularization term not only prevents overfitting but also helps the non-negativity constraint to work properly and supports the convergence of the algorithm. To force non-negativity, each x_i^γ is watched during the iterative process and set to zero whenever their current value is negative. The algorithm is stopped whenever the value of error per pixel is lower than a threshold. Because the error per pixel might never reach that small value for reasons other than illumination (e.g., the 3D shape is a suboptimal estimation), the algorithm stops when a certain number of iterations is reached. Each resulting $\vec{x_i}$ is used as R, G and B values of the i-th directional light, which has been used to render the corresponding illumination cone image $\vec{\Psi}_i$.

The Newton-Raphson update function demands calculation of the first and second derivatives of the summation of Eq. 2 and Eq. 4. The first derivatives are calculated in every iteration. In case of the second derivatives, the result is a Hessian matrix which needs to be inverted. Instead of inverting the Hessian with Singular Value Decomposition (SVD), which is usual for these problems (because of singularity of the matrix a direct inverse is not possible), only the inverse of the diagonal values is calculated and the rest of the Hessian-inverse is set to zero.

Using a synthetic input image, [27] proves that, whenever other rendering parameters are accurately available, this state of the art algorithm converges the sum of least squares to zero.

4. Compressive Sensing for Inverse Lighting

This section describes our approach for inverse lighting of faces using Compressive Sensing (CS) in detail. First we unify the terminology of lighting estimation and CS, we comment on the sensing step and finally we present our algorithm for sparse recovery of the illumination sources.

CS requires a linear sensing model, as indicated in Eq. 5, where $\vec{Y} \in \mathbb{R}^m$ is the vector of measurements, $\boldsymbol{A} \in$

$\mathbb{R}^{m \times n_{atoms}}$ the so-called measurement matrix and $\vec{X} \in \mathbb{R}^{n_{atoms}}$ the sparse vector of coefficients, being n_{atoms} the number of atoms in the sparsity dictionary.

$$\vec{Y} = \boldsymbol{A}\vec{X} \tag{5}$$

The measurements are obtained applying a sensing matrix, $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$ to the dense signal, $\vec{I} \in \mathbb{R}^n$, i.e., $\vec{Y} = \boldsymbol{\Phi}\vec{I}$. Note that the matrix $\boldsymbol{\Phi}$ maps the high-dimensional signal space to a low-dimensional space, $\boldsymbol{\Phi} : \mathbb{R}^n \to \mathbb{R}^m, m \ll n$. If we denote the dictionary as $\boldsymbol{\Psi} \in \mathbb{M}_{n \times n_{atoms}}$ by columns, the signal $\vec{I}$ can be represented as a linear combination of few columns of $\boldsymbol{\Psi}$, i.e., $\vec{I} = \boldsymbol{\Psi}\vec{X}$ and, consequently, the measurement matrix is constructed as $\boldsymbol{A} = \boldsymbol{\Phi}\boldsymbol{\Psi}$. In our case, $\boldsymbol{\Psi}$ is the matrix of cone images introduced in Eq. 1,which allows a sparse representation of the dense vector of face pixels, $\vec{I}$. Both in Eq. 1 and Eq. 5, $\vec{X}$ is the sparse vector containing the intensity values of each possible light source.

We consider now a slightly different notation and stack the channels of the vectorized image by columns, turning the vectors into matrices: $\boldsymbol{Y} = [\vec{Y^\gamma}]_{\gamma \in \Gamma}$, $\boldsymbol{X} = [\vec{X^\gamma}]_{\gamma \in \Gamma}$, being Γ the set of image channels. In this paper we only consider RGB images, i.e., $\Gamma = \{R, G, B\}$, but the method applies to hyperspectral data if a hyperspectral dictionary is provided. Note that both $\boldsymbol{\Psi}$ and $\boldsymbol{A}$ are now 3-dimensional matrices, where the third dimension is $n_{ch} = |\Gamma|$, since a different matrix $\boldsymbol{\Psi^\gamma}$ per channel exists. Each measurement vector $\vec{Y^\gamma}$ is obtained projecting the corresponding channel of the image $\vec{I^\gamma}$ through $\boldsymbol{\Phi}$, i.e., $\vec{Y^\gamma} = \boldsymbol{\Phi}\vec{I^\gamma}$. As introduced in Section 3, a color correction is needed to adapt the color from the 3DMM to the input image. For simplicity, instead of including such correction in our linear model, we undo it in the input image, which is equivalent and always possible, since $\mathrm{rank}(\boldsymbol{T}) = n_{ch} = 3$. If we stack the image channels by rows, we get a matrix $\boldsymbol{I} \in \mathbb{R}^{n_{ch} \times n}$, where, from now on, n is the number of pixels per channel. Then, the measurements are performed on the images given by rows in Eq. 6, where $\mathbf{T}$ is defined in Eq. 3 and $\boldsymbol{O} = \vec{o}\vec{1}_n^{\perp}$.

$$[\vec{I^\gamma}^\top]_{\gamma \in \Gamma} = \boldsymbol{T}^{-1}(\boldsymbol{I} - \boldsymbol{O}) \tag{6}$$

A single sensing matrix, $\boldsymbol{\Phi}$, is used for all channels. The performance of most common types of sensing matrices was evaluated with appropriate data, including pure random and pseudorandom, dense and binary matrices. We selected a Gaussian matrix, with entries i.i.d. drawn from a zero-mean, unit-variance Gaussian distribution. A Bernoulli distribution would equally be valid, as introduced in Section 2.2. The results presented in this paper were also generated using pseudorandom Hadamard-derived binary sensing matrices, with entries $[-1, 1]$ and $[0, 1]$, leading in both cases to results close to those obtained using a Gaussian sensing matrix. Consequently, we can rewrite Eq. 5 as follows:

$$[\vec{Y^\gamma}]_{\gamma \in \Gamma} = [\boldsymbol{A^\gamma}\vec{X^\gamma}]_{\gamma \in \Gamma} \tag{7}$$

where, $\boldsymbol{A^\gamma}$ denotes the 2D matrix obtained from the 3D $\boldsymbol{A}$ matrix for the channel γ, being $\boldsymbol{A^\gamma} = \boldsymbol{\Phi}\boldsymbol{\Psi^\gamma}$. Note that this is close to a Multiple Measurement Vector (MMV) formulation, but not equivalent, since in general $A^{\gamma_i} \neq A^{\gamma_j}, \forall \gamma_i, \gamma_j \in \Gamma, \gamma_i \neq \gamma_j$. Nevertheless, the formulation in Eq. 7 allows us to integrate crucial *a priori* information on the sparsity of the signal, in a MMV manner. Such information is that the matrix $\boldsymbol{X}$ is expected to be k joint sparse, following the definition of joint sparsity given in [11], where $k = |\operatorname{supp}(\vec{X^\gamma})|, \forall \gamma \in \Gamma$. Handling the joint sparsity in our algorithm allows accounting for the physically plausible fact that real illumination sources rarely emit monochromatic light of wavelength coincident to that of one single channel, i.e., if one light source is active, it is active in all channels. This way, we increase the robustness of the estimation and promote sparsity, at the same time we avoid unrealistic lighting configurations.

The algorithm we present here builds upon the ORMP algorithm [23], extending it to handle an arbitrary number of channels. In contrast to existing MMV extensions, we keep a different measurement matrix per channel, but we jointly estimate the index of the dictionary atom to be included in the temporal support at each iteration. For clarity, the pseudo-code for this ORMP for Joint Sparse Inverse Lighting (JSIL-ORMP) is given in Algorithm 1.

Algorithm 1 Order Recursive Matching Pursuit for Joint Sparse Inverse Lighting (JSIL-ORMP)

Initialize: $\boldsymbol{R}^{(0)} = \boldsymbol{Y}, \boldsymbol{X}^{(0)} = \boldsymbol{0}, \Omega^0 = \emptyset, k = 0$

1: **while** $(\|\boldsymbol{R}^{(k)}\|_2 > \varepsilon_{tol})$**and**$(\Delta_{norm}\|\boldsymbol{R}^{(k)}\|_2 > \varepsilon_\Delta)$**and**$(|\Omega^k| < s_{max})$ **do**

2: $k := k + 1$

3: Diag. normalization matrix: $\boldsymbol{N^\gamma}|n^\gamma_{i,i} = \frac{1}{\|\boldsymbol{P}^{\gamma\perp}_{\Omega^k}\vec{A^\gamma_i}\|_2}$

4: $\boldsymbol{G}^{(k)} = [\vec{G^\gamma}]^{(k)}_{\gamma \in \Gamma} = [\boldsymbol{N^\gamma}(\boldsymbol{A^\gamma})^T(\vec{R^\gamma})^{(k-1)}]_{\gamma \in \Gamma}$

5: $i^k = \underset{\substack{i \notin \Omega^{k-1} \\ g_{i,j} > 0, \forall j \in [1, n_{ch}]}}{\arg\max} \|\vec{g_i}\|_2$

6: Update support: $\Omega^k = \Omega^{k-1} \cup i^k$

7: $\boldsymbol{X}^{(k)}_{\Omega^k} = [\vec{X^\gamma}_{\Omega^k}]^{(k)}_{\gamma \in \Gamma} = [\boldsymbol{A^\gamma}^\dagger_{\Omega^k}\vec{Y^\gamma}]_{\gamma \in \Gamma}$

8: Calculate orthogonal projectors: $\boldsymbol{P}^{\gamma\perp}_{\Omega^k} := (\boldsymbol{I}_m - \boldsymbol{A^\gamma}_{\Omega^k}\boldsymbol{A^\gamma}^\dagger_{\Omega^k}),\ \forall \gamma \in \Gamma$

9: Update residual: $\boldsymbol{R}^{(k)} = [\boldsymbol{P}^{\gamma\perp}_{\Omega^k}\vec{Y^\gamma}]_{\gamma \in \Gamma} = [\vec{Y^\gamma} - \boldsymbol{A^\gamma}_{\Omega^k}\boldsymbol{X}^{(k)}_{\Omega^k}]_{\gamma \in \Gamma}$

10: **end while**

where, $\boldsymbol{R}^{(k)}$ is the residual matrix at iteration k and Ω^k the temporal support. In line 4, $\boldsymbol{G}^{(k)} \in \mathbb{R}^{n_{atoms} \times n_{ch}}$ provides, for each channel, the normalized correlations between the current residual vector and the columns of the measurement

matrix. The normalization is provided by the diagonal matrix $\boldsymbol{N}^{\boldsymbol{\gamma}} \in \mathbb{R}^{n_{atoms} \times n_{atoms}}$, whose diagonal elements are $n^{\gamma}_{i,i} = \frac{1}{\|\boldsymbol{P}^{\gamma\perp}_{\Omega^k} \vec{A^{\gamma}_i}\|_2}$, being $\vec{A^{\gamma}_i}$ the ith column of $\boldsymbol{A}^{\boldsymbol{\gamma}}$ and $\boldsymbol{P}^{\gamma\perp}_{\Omega^k}$ the orthogonal projection operator for the channel γ. The selection of the new support element to be added to Ω^k is carried out in line 5. Note the natural handling of the non-negativity of light, achieved by simply requiring that all the elements of the ith row of $\boldsymbol{G}^{(k)}$, $\vec{G_i} \in \mathbb{R}^{n_{ch}}$, are positive, i.e., $g_{i,j} > 0, \forall j \in [1, n_{ch}]$. This automatically avoids selecting dictionary atoms that are negatively correlated with the residual for some of the channels and assures that no negative values will appear in the coefficient vectors, i.e., $x_{i,j} \geq 0, \forall i, j \in [1, n_{atoms}] \times [1, n_{ch}]$. In line 8, $\boldsymbol{I}_m$ denotes the identity matrix of size m and $\boldsymbol{A}^{\boldsymbol{\gamma}\dagger}_{\Omega^k}$) is the Moore-Penrose pseudoinverse of $\boldsymbol{A}^{\boldsymbol{\gamma}}{}_{\Omega^k}$.

The following stopping conditions are contemplated:

1. The residual norm is lower than a threshold, ε_{tol}.
2. The normalized variation of residual norm, $\Delta_{norm}\|\boldsymbol{R}^{(k)}\|_2 = \frac{(\|\boldsymbol{R}^{(k-1)}\|_2 - \|\boldsymbol{R}^{(k)}\|_2)}{\|\boldsymbol{R}^{(k-1)}\|_2}$, is lower than a threshold, ε_{Δ}.
3. The cardinality of the temporal support, $|\Omega^k|$, reaches the maximum sparsity, s_{max}.

If the algorithm ends due to a negative $\Delta_{norm}\|\boldsymbol{R}^{(k)}\|_2$, $\boldsymbol{X}^{(k-1)}_{\Omega^{k-1}}$ is delivered as solution, in order to avoid a degradation of the result due to a wrong last support element. Otherwise, the output of the algorithm is always $\boldsymbol{X}^{(k)}_{\Omega^k}$. The method is robust to variations in the thresholds ε_{tol}, ε_{Δ}. In our case of study, we have observed that even in the most conservative scenario of setting both to zero, the sparsity of the signal is not spoiled. One more case in which the algorithm stops is when no candidate for new support member is found in line 5. This case was found to be uncommon in practice and omitted in Algorithm 1.

5. Results

In this section we evaluate our approach and put it in contrast to that of [27], taken as reference. For both methods the same rendering parameters, i.e., 3D shape, reflectance, texture and rendering method, are identical. In the cases of 100 cone images, the same illumination cone was used by our method and the method in [27]. Our approach is evaluated in four different cases, using 100 and 10000 cone images, with and without an initial face contour-preserving blurring step. In the following, we refer to these cases as CS-100, CS-100-blurred, CS-10000 and CS-10000-blurred. The initial step of Gaussian blur is used in [27] to remove texture that cannot be recovered from the cone images and provides an initial dimensionality reduction. Although included for comparison, a blurring srep is not required in our approach.

In all CS cases, $m = 1000$ measurements are performed, which, in non-blurred images, means a dimensionality reduction of more than 98%. We set the maximum sparsity to $s_{max} = 20$, although this limit was never reached, even when using $n_{atoms} = 10000$ cone images. The tolerance on the l_2 norm of the residual is set to $\varepsilon_{tol} = 0$, pursuing maximum accuracy, and the threshold on residual norm reduction to $\varepsilon_{\Delta} = 1e{-}3$, which is, in practice, very conservative. Our alogorithm was executed twice per experimental case, with different random sensing matrices, to check for stability. Variations in Root Mean Square Error (RMSE) were found to be negligible and the best result was chosen.

5.1. Evaluation on real data

We use six different face images, taken under both real life and controlled lighting conditions, as input data to demonstrate the performance of our method. In each input image, different kinds of illumination effects are present. The results are given in Figure 1. Image 1 is from [28]. In image 1 the illumination is ambient light, still some specular highlights are visible on the forehead, nose and cheek. Image 2 was taken under controlled lighting conditions and shows low saturation and low frequency multilateral illumination. In image 3, blueish specular highlights on the cheek on the left side of the image and some low intensity highlights under the chin are observable. The nose throws a dominant cast shadow on the face. In image 4 and 5, both low and high frequency illuminations are present simultaneously on different sides of the face. In image 4, there is a strong low-frequency illumination on the right side, while on the left side there is an area of low-intensity and high-frequency illumination. In image 5, the dominant illumination areas on the left and right side of the face are of different frequency and colors. Image 6 shows specular highlights on several face areas. There is a cast shadow under the nose and an attached shadow on the right side of the image, under the chin and the cheek. Note the colorful light within the attached shadow under the chin.

In general, the case CS-100 appears to deliver more reliable results in all experiments. Because our method promotes sparsity, it reconstructs the illumination setup with lower number of lights. As a downside, this might lead to ghost shadows or too bright highlights. On the positive side, lower number of lights allows for shorter rendering times and more realistic reconstruction of extreme specular highlights and cast shadows. Distributions of light sources, showing color and location around the face are given in Figure 2 for two cases, for reference and proposed method. For image 6 (last row) a sparsity enhancement of more than 80% is achieved, at negligible cost in terms of RMSE.

In Table 1 we provide the RMSEs between the images rendered using the estimated lighting and the originals. We use the l_2 norm only for allowing comparison between algo-

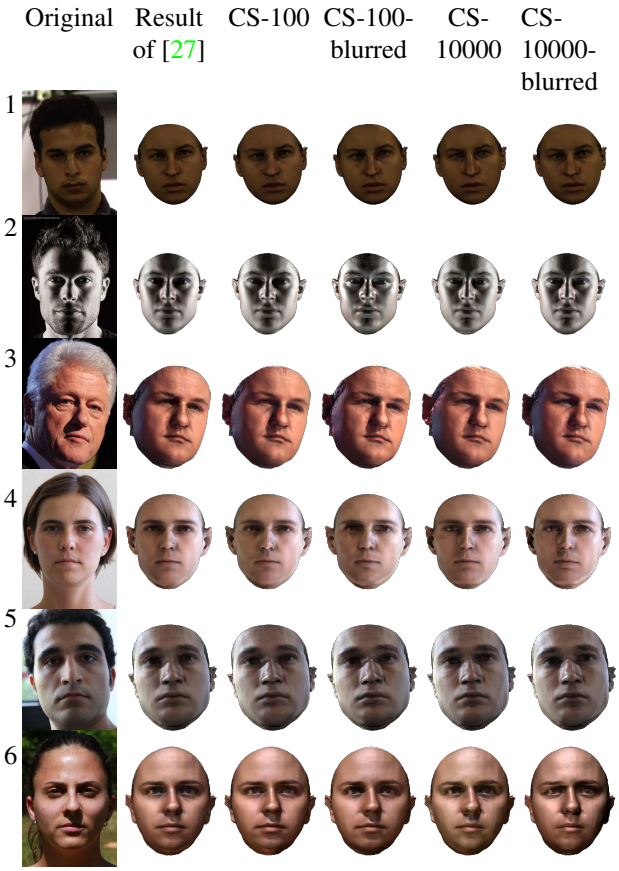

Figure 1. Experimental results for the six different real images of faces considered in this paper. The reconstruction results are organized by columns, from left to right, as follows: the first column shows the original input image, the second shows the result obtained using the method presented in [27] and le last four columns present the results of the proposed algorithm in four different recovery scenarios: Using $n_{atoms} = \{100, 10000\}$ cone images, both with and without applying an initial step of *ad hoc* Gaussian blur and decimation. Input image 2 is courtesy of ©Barrie Spence 2013

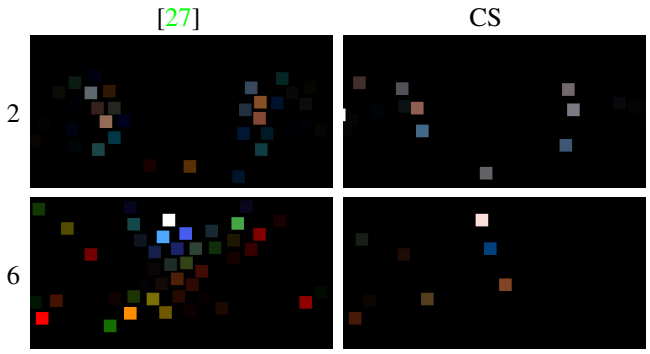

Figure 2. Map of active light sources obtained with the method in [27] (left column) and with the proposed method (right column), for the input images 2 and 6. The results of our method are using 10000 cone images in 2 and 100 in 6.

rithms, but note that cannot be considered an optimal quantitative measure for illumination estimation. Note that, even with much fewer light sources, we get RMSEs that are very close to those achieved in [27], still slightly higher, in general.

ID	[27]	CS-100	CS-10000
1	3.065e-2	3.210e-2	3.106e-2
2	1.464e-1	1.582e-1	1.473e-1
3	1.077e-1	1.110e-1	1.037e-1
4	5.767e-2	5.965e-2	6.045e-2
5	7.175e-2	7.396e-2	7.194e-2
6	1.012e-1	1.059e-1	1.047e-1

Table 1. Root Mean Square Error (RMSE) between the image rendered using the estimated lighting and the original image. Images are rendered with average face texture to focus on the effects of illumination. The first column contains the results obtained applying the method in [27], the second and third columns summarize the results obtained using the proposed method, with 100 and 10000 cone images, respectively.

In terms of execution time, our method has shown to be, at least, twice as fast as that of [27]. Additionally, if runtime is an important parameter, a simpler version of our algorithm building on OMP can be adopted to avoid the orthogonal projection at each iteration. Such modification allows for runtimes at least two orders of magnitude lower than those for the method in [27] using 100 cone images and opens the door for a real time implementation. Moreover, the proposed algorithm does not require parameter tuning.

5.2. Robustness to noise

Methods operating in pixel domain, apart from suffering from overwhelming dimensionalities, are typically sensitive to noise. Using a pseudorandom mapping as dimension-reduction method allows for filtering out random noise, since random variables are poorly correlated between each other. Unfortunately, pseudorandom mappings also tend to decrease the power of the signal and the signal-to-noise ratio may be significantly reduced in the compressed domain. For this reason, an analysis of the performance of the proposed method under a wide range of noise levels becomes necessary.

We consider the addition of zero-mean Gaussian noise to the input color images. In order to allow comparison to ground truth, we use the realistic illumination recovered from the images in Section 5.1 to generate six synthetic face images. For each image, six levels of noise are considered, yielding SNRs from $40\,\mathrm{dB}$ to $0\,\mathrm{dB}$, with $10\,\mathrm{dB}$ step size, plus a noiseless case. The analysis is carried out for the four experimental cases considered, namely CS-100, CS-100-blurred, CS-10000 and CS-10000-blurred. Very similar values and exactly the same trend was observed for the

four cases. Figure 3 shows the RMSE between the image rendered using the recovered lighting and the original noiseless image, for different SNRs of the input image, in the case of CS-100. The RMSEs obtained for the noiseless images are very close to those obtained for the 40 dB cases and are omitted. Figure 4 illustrates the quality of the recovered lighting from noisy input images for Image 3 in the CS-100 case (red line in Figure 3). Note the faithful recovery of the illumination for SNRs of the input image as low as 10 dB, as well as the correct estimation of the main light directions in the 0 dB case, thanks to our robust joint-support estimation. Provided that random noise is not correlated between channels, estimating the sparse joint support in a multichannel manner increases the robustness to noise.

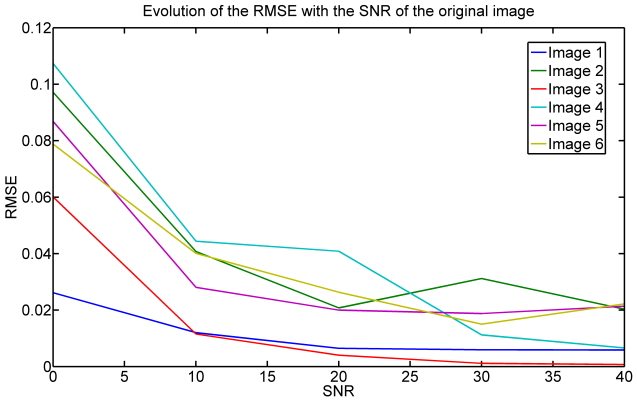

Figure 3. Root Mean Square Error (RMSE) between the image rendered using the estimated lighting and the original noiseless image, for different noise levels in the input image. The results were obtained using 100 cone images, without initial blurring step (CS-100). The plots show very accurate recovery for SNRs as low as 10 dB in all cases.

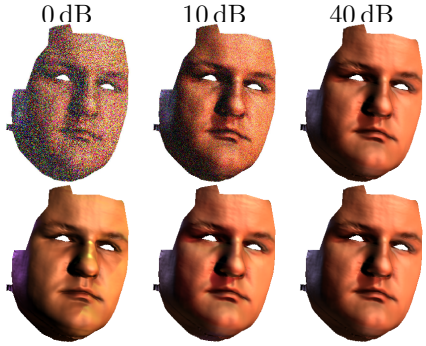

Figure 4. Recovered illuminations from noisy input images for the Image 3, using CS100. The first row shows the input images, corrupted with different levels of zero-mean Gaussian noise, showing SNRs of 0, 10 and 40 dB. The second row shows the recovered illumination from the noisy images.

5.3. Dependence of error on compression rate

All results summarized in Figure 1 were obtained using the same number of measurements, $m = 1000$. In order to study the stability of our method with the number of measurements, we repeat the experiments for different values of $m \in [50, 1000]$, with step size of 50. This covers compression rates between 98.2% and 99.9% in the non-blurred cases. Figure 5 shows the RMSE between the image rendered using the estimated lighting and the original image, for each value of m considered, in the case of CS-100. Note the higher error obtained for excessively low values of m and the high stability almost all along the considered range.

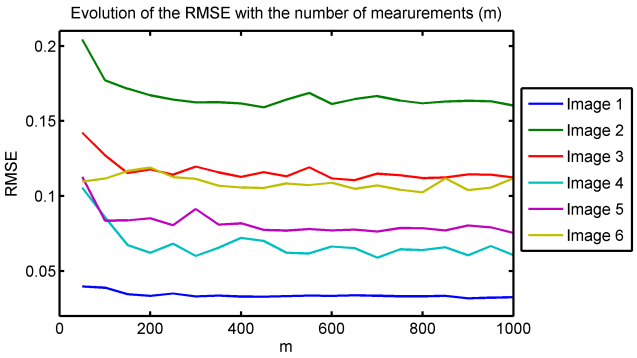

Figure 5. Root Mean Square Error (RMSE) between the image rendered using the estimated lighting and the original image, for different number of measurements, $m \in [50, 1000]$. The results were obtained using 100 cone images, without initial blurring step (CS-100). The plots show stable recovery for compression rates higher than 99% in all cases.

6. Conclusions and Future Work

In this paper, we have presented a novel approach for inverse lighting of faces using CS. Our recovery algorithm is a *greedy* method based on the ORMP [23], which naturally handles the non-negativity of light sources in multichannel images. A joint support selection schema provides enhanced robustness to uncorrelated noise.

Experimental evaluation with challenging real images shows that our method is able to provide a much sparser illumination setup than previous methods also based on an illumination cone, often one order of magnitude sparser, while achieving equivalent performance in terms of RMSE. The dimensionality reduction provided by the stochastic sensing, higher than 98% for the images in this paper, allows reduced runtimes, handling large input images and a high number of cone images.

References

[1] O. Aldrian and W. A. P. Smith. Inverse rendering of faces with a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(5):1080–1093, May 2013. 2

[2] R. Baraniuk. Compressive sensing [lecture notes]. *Signal Processing Magazine, IEEE*, 24(4):118–121, July 2007. 3

[3] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible illumination conditions? *Int. Journal of Computer Vision*, 28(3):245–260, 1998. 1, 2

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 1, 2, 4

[5] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, Feb 2006. 2, 3

[6] E. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, Dec 2006. 3

[7] E. Candès and M. Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, March 2008. 3

[8] E. J. Candès. Compressive sampling. In *Proceedings oh the International Congress of Mathematicians*, pages 1433–1452, August 2006. 3

[9] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, Aug. 2006. 3

[10] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM JOURNAL ON SCIENTIFIC COMPUTING*, 20:33–61, 1998. 3

[11] M. Davies and Y. Eldar. Rank awareness in joint sparse recovery. *Information Theory, IEEE Transactions on*, 58(2):1135–1146, Feb 2012. 5

[12] P. Debevec. A median cut algorithm for light probe sampling. In *ACM SIGGRAPH 2006 Courses*, SIGGRAPH '06, New York, NY, USA, 2006. ACM. 2

[13] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. *Proc. 27th Annu. Conf. Comput. Graph. Interact. Tech. - SIGGRAPH '00*, pages 145–156, 2000. 2

[14] R. DeVore and V. Temlyakov. Some remarks on greedy algorithms. *Advances in Computational Mathematics*, 5(1):173–187, 1996. 3

[15] D. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, April 2006. 2, 3

[16] Y. Eldar and G. Kutyniok. *Compressed Sensing: Theory and Applications*. Compressed Sensing: Theory and Applications. Cambridge University Press, 2012. 3

[17] M. Fuchs, V. Blanz, and H.-P. Seidel. Bayesian relighting. In *Proceedings of the Sixteenth Eurographics Conference on Rendering Techniques*, EGSR'05, pages 157–164, Aire-la-Ville, Switzerland, 2005. Eurographics Association. 2

[18] P. Graham, B. Tunwattanapong, J. Busch, X. Yu, A. Jones, P. Debevec, and A. Ghosh. Measurement-Based Synthesis of Facial Microgeometry. *Comput. Graph. Forum*, 32(2pt3):335–344, May 2013. 2

[19] C. Li, K. Zhou, and S. Lin. Intrinsic face image decomposition with human face priors. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 218–233. Springer, 2014. 2

[20] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, Dec 1993. 3

[21] X. Mei, H. Ling, and D. Jacobs. Illumination recovery from image with cast shadows via sparse representation. *Image Processing, IEEE Transactions on*, 20(8):2366–2377, Aug 2011. 3

[22] S. Muthukrishnan. Data streams: Algorithms and applications. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, pages 413–413, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics. 3

[23] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, Apr. 1995. 3, 5, 8

[24] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, Twenty-Seventh Asilomar Conference on*, pages 40–44 vol.1, Nov 1993. 3

[25] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 117–128, New York, NY, USA, 2001. ACM. 1

[26] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pages 986–993, Washington, DC, USA, 2005. IEEE Computer Society. 2

[27] D. Shahlaei and V. Blanz. Realistic inverse lighting from a single 2d image of a face, taken under unknown and complex lighting. In *Proceedings of the 11th IEEE International Conference on Automatic Face and Gesture Recognition*, FGR '15, 2015. 1, 2, 4, 6, 7

[28] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, FGR '02, pages 53–, Washington, DC, USA, 2002. IEEE Computer Society. 6

[29] T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, J. McAndless, J. Lee, A. Ngan, H. W. Jensen, and M. Gross. Analysis of human faces using a measurement-based skin reflectance model. *ACM Trans. Graph.*, 25(3):1013–1024, July 2006. 1, 2

[30] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, Feb 2009. 4

[31] L. Zhuang, T.-H. Chan, A. Yang, S. Sastry, and Y. Ma. Sparse illumination learning and transfer for single-sample face recognition with image corruption and misalignment. *Int. Journal of Computer Vision*, pages 1–16, 2014. 3, 4