Pattern Recognition 90 (2019) 308–324

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

SSR2: Sparse signal recovery for single-image super-resolution on faces with extreme low resolutions

Ramzi Abiantun*, Felix Juefei-Xu*, Utsav Prabhu, Marios Savvides

CyLab Biometrics Center, Electrical and Computer Engineering Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Received 22 August 2018
Revised 8 November 2018
Accepted 25 January 2019
Available online 28 January 2019

Keywords:
Sparse signal recovery (SSR)
Single-image super-resolution (SSR)
Extreme low resolution

ABSTRACT

Automatic face recognition in the wild still suffers from low-quality, low resolution, noisy, and occluded input images that can severely impact identification accuracy. In this paper, we present a novel technique to enhance the quality of such extreme low-resolution face images beyond the current state of the art. We model the correlation between high and low resolution faces in a multi-resolution pyramid and show that we can recover the original structure of an un-seen extreme low-resolution face image. By exploiting domain knowledge of the structure of the input signal and using sparse recovery optimization algorithms, we can recover a consistent sparse representation of the extreme low-resolution signal. The proposed super-resolution method is robust to noise and face alignment, and can handle extreme low-resolution faces up to 16x magnification factor with just 7 pixels between the eyes. Moreover, the formulation of the proposed algorithm allows for simultaneous occlusion removal capability, a desirable property that other super-resolution algorithms do not possess, to the best of our knowledge. Most importantly, we show that our method generalizes on real-world low-quality surveillance images, showing the potentially big impact this can have in a real-world scenario.

1. Introduction

During the 2013 Boston Marathon on April 15, two pressure cooker bombs exploded at 2:49 p.m. EDT near the finish line, killing 3 spectators and injuring 264 others. Shortly after, the FBI personnel investigating the scene acquired footage of the explosion [1] and the preceding events, including placement of the bombs, from street surveillance cameras. However, due to the grainy, *low-resolution (LR)* quality of the images obtained, current commercial face recognition software was unable to recognize the primary suspects involved in the bombing, and the authorities had to subsequently resort to crowdsourcing to obtain higher quality images of the suspects from other sources such as cell phone photos. This crowdsourcing resulted in higher resolution images on April 19, which could have been processed by commercial face recognition systems [2]. However, the initial low-resolution images (Fig. 1(a)) were unsuitable for this task. The incident highlighted the importance of image quality towards real-world face recognition and the need for the development of algorithms to overcome the deterioration caused by unconstrained LR input footage.

Super-resolution (SR) refers to the resolution enhancement of the visual information contained in a low-resolution image. Low-resolution images are the result of how far the object of interest is from the camera, the specifications of the imaging sensor (and its tolerance to noise and low-light sensitivity), and the quality of the optical lenses attached to the sensor (represented by the modulation transfer function (MTF)). The widespread availability of affordable digital imaging devices, such as cell phone cameras, surveillance cameras, *etc*., comes as a result of the recent mass production and the shrinking in size of these devices. However, this does not always translate into increased image quality, and typically the visual quality of footage obtained by these low-cost devices remains poor.

The holy grail of face recognition is to build a system that can handle real-world data. This data emerges from surveillance footage (as in the case of the Boston Marathon incident), or mobile devices used at a distance, where it is not only the viewpoint that poses a challenge but also the low-resolution image. Even though face recognition algorithms do not require megapixel images, having *high-resolution (HR)* texture information goes a long way towards improving recognition performance [3]. Most algorithms can produce respectable results with image resolution as low as 50×50 pixels [4]. The problem is that most faces in the wild do not even meet these minimum size requirements. For instance, most surveillance cameras currently deployed consist of a wide-angle lens (to capture as much of the scene as possible) coupled to a 720 p HD (1280×720) pixel sensor. A subject standing

* Corresponding authors.
E-mail address: juefei.xu@gmail.com (F. Juefei-Xu).

https://doi.org/10.1016/j.patcog.2019.01.032
0031-3203/

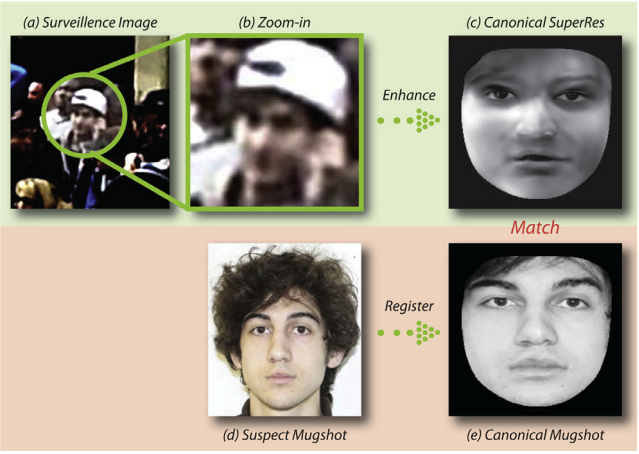


Fig. 1. (a) Surveillance image of the Boston Marathon bombing suspect, (b) zoom and crop of (a) showing lack of visual details for recognition, (c) our super-resolution of the same face displayed in a canonical format, (d) a high-resolution image of the same suspect, (e) the same high-resolution image in the canonical format for matching.

a few meters away will not reproduce a face with a significant number of pixels between the eyes, rendering it impossible even for a human operator to identify. In one of the earlier studies on this same topic, Bachmann [5] observes that there is an abrupt fall in identification efficiency by human operators when the faces become smaller than 24×18 pixels.

Super-resolution, which aims to reverse this last problem, is a severely ill-posed problem, where the solution does not always exist, and when it exists, it is not unique and is very sensitive to perturbation in the input. Despite years of active research, it still remains an open challenge, particularly when trying to break the 4x magnification barrier in resolution. In this paper, we present a simple yet novel technique to attempt to solve the low-resolution problem. We rely on sparse feature recovery and a multi-resolution face model to develop a *single-image* super-resolution (or face hallucination) technique. If we can achieve great results using single image resolution, we can enhance this solution using a sequence of images.

One important fact worth mentioning is that degradations such as extremely low-resolution and facial occlusions usually come in a bundle. To the best of our knowledge, the current state-of-the-art super-resolution algorithms can not and do not deal with occlusion removal. They usually rely on another occlusion removal module should they choose to do so. Our proposed method, however, can simultaneously deal with the super-resolution as well the occlusion removal tasks under the same framework, a very desirable property that other super-resolution algorithms do not possess. This multi-tasking capability is possible due to the formulation of our algorithm that treats both super-resolution and occlusion removal as a missing data recovery challenge.

This paper is structured as follows: In Section 2, we first briefly overview some of the classical approaches to face super-resolution. In Section 3, we develop our novel sparsity-based approach towards face super-resolution, which involves learning a face model based on a multi-resolution pyramid of faces. In Section 4, we evaluate our method on synthetic low-resolution images (obtained from downsampling the high-resolution images). We also show that our method is inherently robust to noise by reformulating it as a Bayesian approach. Next, we show that our super-resolution algorithm has the occlusion removal capability, in a multi-task face recovery experiment. In Section 5, we demonstrate the real-world applicability of the technique by showing its effectiveness on the Boston Marathon bombing case. Finally, we conclude our work in Section 6.

2. Related work

2.1. General image super-resolution

Most single-image super-resolution techniques are limited because they are constrained by the number of available pixels. Early methods mostly relied on interpolation techniques, such as nearest-neighbor, bilinear or cubic B-spline interpolation [6] or convolution kernels [7]. Interpolation-based methods assume global continuity and maintain smoothness constraints that often produce results with blurred edges and textures which are essentially unusable in a face recognition framework. Edge-preserving interpolation techniques have been proposed, such as adaptive splines [8], and POCS (projection onto convex sets) interpolation [9]. Interpolation techniques aware of edge information have also been proposed. Nonlinear interpolation with edge fitting [10] incorporates local edge fitting to avoid interpolation across edges. Huang et al. [11] introduce a self-exemplar based SR method by searching recurring patterns within and across scales of the same image. By further allowing affine and projective transformations on the patterns, the proposed method has reached the state-of-the-art results among all the self-exemplar based SR methods. Wang et al. [12] propose an ensemble based sparse coding network for single-image super-resolution. Zhang et al. [13] attempt to fuse the brain atlas from the high-thickness diagnostic MR images that are prevalent for clinical routines. By incorporating a novel super-resolution strategy, as an extension of the conventional group-wise registration, the required atlas can be better constructed.

2.2. Face image super-resolution

Super-resolution of face images is slightly different than super-resolution of general images because we can make more assumptions about the structure of the human face, and have face-specific image priors. In their seminal work on super-resolution, Baker and Kanade [3] abandon the Markov random fields (MRF) framework for a more general Bayesian maximum *a posteriori* (MAP) formulation which is more suitable for synthesizing global textures as with face images. They use a large number of training images to compute a multi-resolution pyramid of features (such as Laplacian and gradients features). Each level of the pyramid corresponds to a different resolution that is obtained by reducing the original native full resolution. In testing, given an input low-resolution image, they populate the top of the pyramid of features (that they call “the Parent Structure”), and for every pixel location, they exhaustively search (using the nearest-neighbor approach or gradient descent) the training set for a pixel value that generates a similar feature vector. Moreover, they use a Bayesian framework to incorporate more than one input low-resolution images. Assuming the input images have been accurately aligned together (using control points on the face or a registration method such as optical flow), the nearest-neighbor search is replaced by a MAP computation that reflects the likelihood of observing several low-resolution pixel values and a prior on the high-resolution values.

Liu et al. [14] combine global parametric and local non-parametric models in a two-step statistical approach. They assume a high and low frequency mixture model, and for super-resolution, the reconstructed low-frequency information is inferred by a global linear model that learns the relationship between high- and low-resolution images, while the reconstructed high-frequency information is captured by learning the residual between the original high-resolution and the super-resolution image using a patch-based nonparametric Markov network. In [15], a simple

principal component analysis (PCA)-based global approach has been attempted. Given high-resolution training images, the authors downsample them and build a PCA subspace of reduced-resolution images, where they project the input low-resolution image to obtain the PCA coefficients. The authors then use those low-resolution induced coefficients to reconstruct the high-resolution equivalent face using the full-resolution eigenfaces. To make sure that the final reconstruction is "face-like", they introduce artificial constraints on the coefficient values. These constraints are a function of the eigenvalues learned from the training of every principal component.

Yang et al. [16] exploit the properties of sparse image representation for single-image super-resolution. Their patch-based local approach simultaneously learns two distinct over-complete dictionaries: one for high-resolution patches, and one for low-resolution ones. Given an input low-resolution patch, its sparse representation in the low-resolution dictionary is used to recover a high-resolution patch from the high-resolution dictionary. Local consistency is achieved by requiring the patches to overlap and requiring the high-resolution patches to agree on the overlapped area. In the specific case of face hallucination, a two-step approach is adopted, similar to previous methods. The first step is a global approach to reconstruct what they call a "medium high-resolution" smooth face using non-negative matrix factorization (NMF) [17], which is then enhanced with high-frequency information using the local patch-based approach that makes use of the sparse feature extraction using the coupled dictionaries learned at the training stage. Another work [18] also exploits sparsity. The basic idea in this work is to use kernel ridge regression to learn the mapping between high- and low-resolution images. To avoid blurring artifacts, a post-processing step that relies on a prior model is employed.

Ma et al. [19] hallucinate the high-resolution image patch using the same position image patches of each training image. The optimal weights of the training image position patches are estimated and the hallucinated patches are reconstructed using the same weights. The final high-resolution face image is obtained by integrating the hallucinated patches.

Gao et al. [20] have proposed a locality-constrained double low-rank representation (LCDLRR) method to improve upon position patch based face hallucination. LCDLRR tries to directly use the image-matrix based regression model to compute the representation coefficients to maintain face structural information. A low-rank constraint is imposed on the representation coefficients to adaptively select the training samples that belong to the same subspace as the inputs.

Jiang et al. [21] also try to improve upon the position patch based face super-resolution approach by incorporating a locality constraint into the least square inversion problem to maintain locality and sparsity simultaneously. This method is termed locality-constrained representation (LcR).

Jiang et al. [22] have proposed a coupled-layer neighbor embedding (CLNE) method which contains the LR and HR layer. The LR layer models the local geometrical structure of the LR patch manifold which is characterized by the reconstruction weights of the LR patches. The HR layer is the intrinsic geometry that can constrain the reconstruction weights geometrically. CLNE can achieve a robust neighbor embedding by iteratively updating the LR patch reconstruction weights and the estimated HR patch.

Huang and Wu [23] study the face image super-resolution problem under resource-limited environment. Their method utilizes multiple local linear transformations (LLT) to approximate the nonlinear mapping between LR and HR images in the pixel domain. The affine transformations between LR and HR face patches are estimated from training examples and the LLT-based reconstruction is achieved by applying the transformations to all patches of an LR input image, followed by a refinement step using the POCS algorithm.

Zeng et al. [24] expand the training data for improved facial image SR. Three constraints (the local structure constraint, the correspondence constraint, and the similarity constraint) are proposed to generate new training data where local patches are expanded with different parameters.

Huang et al. [25] have applied canonical correlation analysis (CCA) to maximize the correlation between the local neighbor relationship of high and low resolution images. CCA is used separately for reconstructing global face appearance as well as local facial details. An and Bhanu [26] instead propose to use 2D CCA for better preserving the 2D structure of the faces.

Jiang et al. [27] have proposed a sparse representation based noise robust super-resolution approach that incorporates smooth prior to enforce similar training patches having similar sparse coding coefficients. The method fuses LASSO-based smooth constraint and locality-based smooth constraint to the least squares representation-based patch representation for obtaining stable reconstruction weights especially when the noise level of the input LR image is high.

Akyol et al. [28] present a face super-resolution method based on generative models and utilizes both the shape and texture components. The main idea is that the image details can be synthesized by global modeling of accurately aligned local image regions. In order to achieve sufficient accuracy in alignment, shape reconstruction is considered as a separate problem and solved together with texture reconstruction in a coordinated manner.

Nguyen et al. [29] provide a comprehensive review of existing super-resolution approaches for biometric modalities, including face (2D and 3D), iris, gait and latent prints (fingerprint and palmprint) and other emerging modalities.

The aforementioned face image super-resolution techniques do not tackle very challenging low-resolution cases, a moderate 4x downsample factor can be found across the board [19–21,23–29], with the exception of [22] which deals with 8x downsampling. Our work instead takes on extremely low-resolution face images with up to 16x downsample factor.

2.3. Deep learning based super-resolution

More recent advances for image super-resolution include deep learning based approaches such as [30–34]. Dong et al. [30] have a seminal work on image super-resolution using convolutional neural networks, which is termed SRCNN, and is later extended in [31]. This is the very first attempt to tackle the image super-resolution problem in an end-to-end learnable fashion. Kim et al. [32] come up with an improved CNN-based SR solution by going deep (20 convolutional layers), and explicitly modeling the residual image. Their method converges much faster than SRCNN by utilizing adjustable gradient clipping, and residual learning, and achieves better SR results. Another work by the same authors [33] explores recursive filtering in the deep convolutional neural networks. The idea is to re-use the same filter over and over again on the feature maps recursively before it gets updated, which essentially allows many intermediate reconstruction outputs to be fused at the end. In addition, skip connections are added so that each intermediate stage gets a direct supervision signal, which makes the learning more effective and less prone to gradient vanishing.

With the help of generative adversarial network (GAN), the SRGAN [34] approach can achieve more photo-realistic 4x super-resolution. The gist is that instead of minimizing some mean squared loss between the super-resolved image and the ground-truth high-resolution image as commonly practiced, new adversarial loss and content loss are used. The adversarial loss is

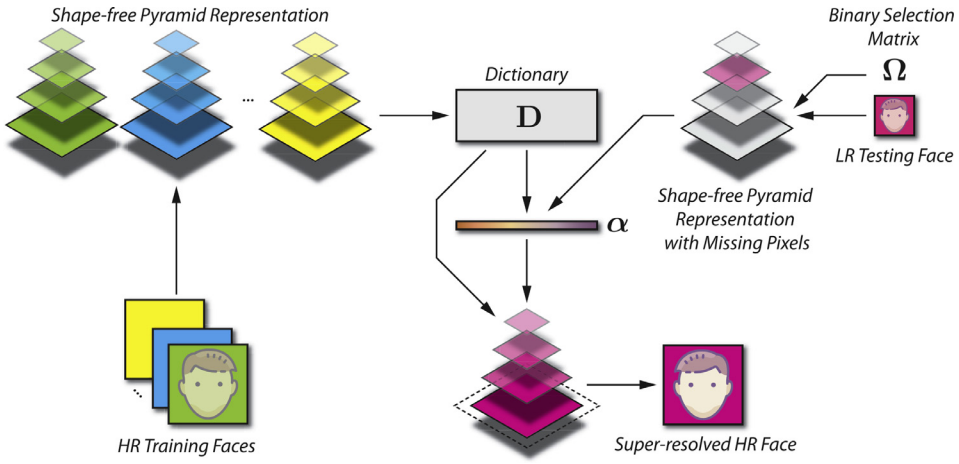


Fig. 2. Overview of the proposed SSR2 method.

attached to a deep CNN based discriminator network to differentiate between the super-resolved images and the original ground-truth high-resolution images and to push the solution to the natural image manifold, which is major merit of GAN-based approaches. The content loss ensures that deep features, say using a VGG-19 network, match between the super-resolved and the ground-truth images.

3. Feature extraction to recover missing resolutions

In this section, we present a simple global technique to achieve face super-resolution by leveraging the ability of sparsity-based methods to solve missing-data problems [35]. The flowchart of the proposed method is shown in Fig. 2. In a nutshell, we first extract shape-free representations from a set of high-resolution training face images and formulate the pyramid representation. Then, these pyramid shape-free representations will be used to train a dictionary for later sparse signal recovery. When a low-resolution face image comes in, the extracted pyramid shape-free representation will have missing pixels at higher resolution hierarchy. Through sparse signal recovery and reconstruction using the full dictionary, we can recover the corresponding high-resolution face image. We will detail the involved components individually.

3.1. Subspace modeling to extract features stable under missing dimensions

Global face super-resolution approaches are criticized for not being able to render sharp edges and high-frequency information, or that PCA-based methods often degenerate towards the mean face. We disprove this common belief by individually detailing the fundamentals of our technique. First, we note that the problem at hand is a texture reconstruction problem. Hence, we need to decouple the shape information from the texture information.

3.1.1. A canonical shape-free representation for accurate texture analysis

Super-resolution is principally a texture-driven problem, hence it would be beneficial for our data representation to contain mostly texture features, isolated from the shape of the face. We define "shape" by the x and y coordinates of a specific set of predefined landmarks on a face. This shape information is eliminated from the data by enforcing all faces to adopt the same "mean" face shape before modeling the subspace, *i.e.*, all facial features, such as eyebrows, eyes, nose, mouth, *etc.*, should have the same dimensions and locations in all faces. To accomplish this, a simple global transformation such as affine is insufficient. We require a local approach, where every triangle bounded by any three control points can be transformed independently.

In [36], a piece-wise linear warping scheme was introduced, which could be used to warp all faces to a particular mean shape. However, such an approach would introduce discontinuities and other artifacts which would negatively impact the subspace modeling of the texture information. We choose to make use of an efficient 3D modeling technique called 3D generic elastic models (3DGEM) [37], and construct a 3D generic structure [38]. We then render all face textures with this common structure to normalize the shape of the faces. A flowchart of this technique is shown in Fig. 3. This results in a completely shape-normalized face image, with much smoother rendering free of high-frequency discontinuity artifacts [35]. In Fig. 4, we depict three frontal faces from the MPIE database [39] with a traditional crop using eye coordinates and their equivalent shape-free representation for comparison.[1]

Once we have transformed all of our training data using this technique, we use the shape-normalized renders to construct a texture-only subspace of the face. Since the purpose of the subspace is to represent the texture feature space accurately (rather than to learn discriminative characteristics), simple representations such as singular value decomposition (SVD) [41], NMF [17], clustering or other dictionary learning approaches such as K-SVD [42] built from several thousands of face images transformed in the shape-free domain are sufficient to capture the principal texture variation. In this paper, our framework allows for SVD-type dictionaries for their sparse representation properties.

3.1.2. Super-resolution reformulated as a missing data challenge

The core of our resolution-enhancing technique consists of modeling the relationship between high- and low-resolution face images, and then, from an unseen low-resolution query face, infer the high-resolution equivalent face. From this perspective, super-resolution translates into missing data problem. To model the correlation between different resolutions, we build a multi-resolution pyramid, similar to [3], and model its subspace. This pyramid will enable us to avoid explicitly estimating the resolution-reduction function or parameters of a point-spread-function. We represent this high-dimensional multi-resolution subspace by a matrix **D** of

[1] We noticed that commercial FR software behave erratically with this representation due to the edge introduced by the blank background [40]. This will be the topic of investigation in a later publication. This observation motivated our approach for obtaining a sparse face representation that can be used for matching.

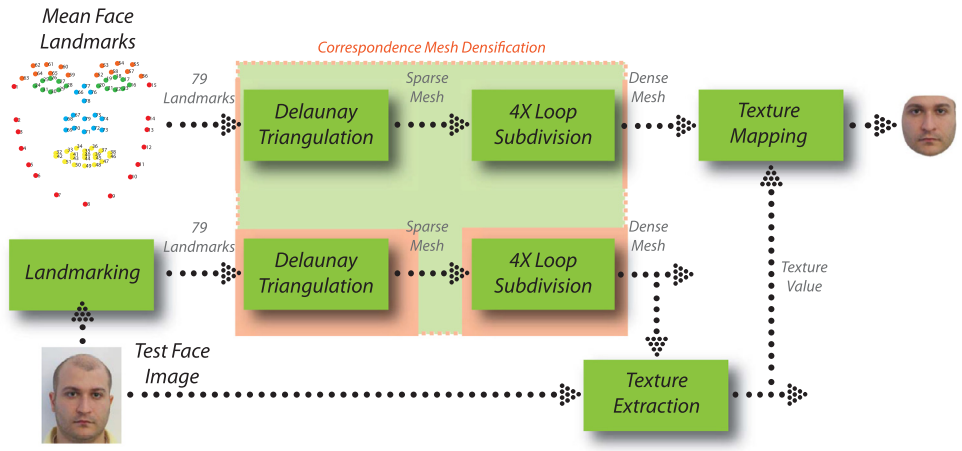


Fig. 3. Canonical shape-free generation module for accurate texture analysis and reconstruction.

Fig. 4. Comparison between traditional face crop and shape-free representation. The latter is essential for feature extraction in missing texture information scenario.

vectorized dictionary atoms and a mean vector $\mathbf{m}$ trained on a Gaussian pyramid [43] of training images as follows:

Given a training image I_i in the canonical shape-free representation, the Gaussian pyramid $G_0(I_i), \ldots, G_k(I_i)$ for such an image is depicted in Fig. 5. Following [43], the bottom level of the pyramid is the image itself, and every subsequent level is obtained by $G_{i+1}(I) = \text{Reduce}(G_i(I))$ where the Reduce operator is defined by the following equation:

$$\text{Reduce}(I)[i,j] = \sum_{m=1}^{5}\sum_{n=1}^{5} w[m,n]I[2i+m, 2j+n] \tag{1}$$

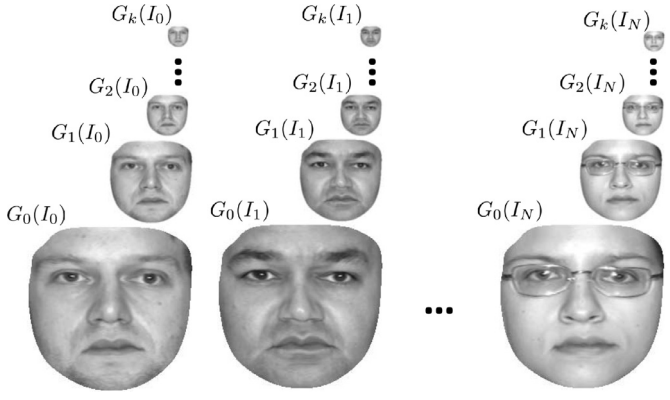


Fig. 5. Depiction of a Gaussian pyramid of $k+1$ levels for $N+1$ face images in the shape-free representation taken from the MPIE dataset.

w, in this case, is a 5×5 low-pass Gaussian filter kernel. We build the training subspace by concatenating the vectorized images of all $k+1$ levels into a single column. For testing, we assume we only have access to the kth level of the pyramid of an unseen face image. The active dimensions in this current "missing data" problem are the dimensions corresponding to the pixels of the input low-resolution image. The missing dimensions are the dimensions corresponding to the pixels of the images in the lower levels of the pyramid. In the case where a medium-resolution image is available (as is common in most super-resolution algorithms that rely on Gaussian pyramids), the smaller levels of the pyramid can be populated, and the active dimensions become the top of the pyramid, while the missing dimensions are at the bottom of the pyramid.

Let $\mathbf{x}'$ be the vector of observed pixels of $\mathbf{x}$, *i.e.*, if $\mathbf{\Omega}$ is a $d' \times d$ binary row selection matrix ($d' < d$), then $\mathbf{x}' = \mathbf{\Omega x}$. We need to solve for the signal representation vector $\boldsymbol{\alpha}$, so in the case of an overdetermined system ($d' > N$), we can minimize the following cost function $J(\boldsymbol{\alpha})$:

$$\mathbf{J}(\boldsymbol{\alpha}) = \left\| \mathbf{x}' - \mathbf{\Omega}(\mathbf{D}\boldsymbol{\alpha} + \mathbf{m}) \right\|_2^2 \tag{2}$$

Setting the gradient $\nabla \mathbf{J}(\boldsymbol{\alpha})$ *w.r.t.* $\boldsymbol{\alpha}$ to zero we obtain:

$$\hat{\boldsymbol{\alpha}}_{\text{lse}} \approx (\mathbf{D}^T\mathbf{\Omega}^T\mathbf{\Omega}\mathbf{D})^{-1}(\mathbf{D}^T\mathbf{\Omega}^T)(\mathbf{x}' - \mathbf{\Omega}\mathbf{m}) \tag{3}$$

As the number of missing dimension increases and $d' < N$, the left pseudo-inverse becomes infeasible. Alternatively, instead of minimizing the least squares of the error vector, as in Eq. (2), we can opt to minimize the following cost function:

$$\text{minimize} J(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_2 \text{ subject to } \mathbf{x}' = \mathbf{\Omega}(\mathbf{D}\boldsymbol{\alpha} + \mathbf{m}) \tag{4}$$

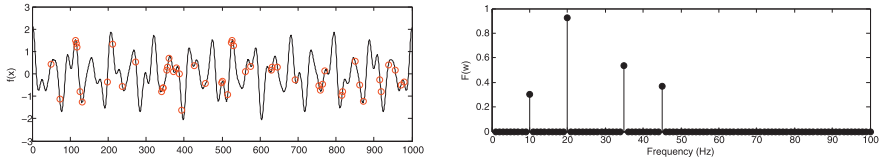
The Euclidean norm allows us to elegantly solve the equation using Lagrange multipliers. Solving for the optimality conditions yields the following answer:

$$\hat{\boldsymbol{\alpha}}_{\text{mn}} \approx (\mathbf{D}^T\mathbf{\Omega}^T)(\mathbf{\Omega}\mathbf{D}\mathbf{D}^T\mathbf{\Omega}^T)^{-1}(\mathbf{x}' - \mathbf{\Omega}\mathbf{m}) \tag{5}$$

The vector $\hat{\boldsymbol{\alpha}}_{\text{mn}}$ represents the *minimum-norm* in ℓ_2. We can assess the quality of the obtained feature vectors by visually inspecting the quality of the reconstruction. Using the coefficient vector we obtained, we can reconstruct the full pyramid and extract whichever level of the pyramid we want following $\hat{\mathbf{x}}_{\text{mn}} = \mathbf{D}\hat{\boldsymbol{\alpha}}_{\text{mn}} + \mathbf{m}$.

An alternative formulation to Eq. (5) is to consider the ℓ_1 norm instead of the ℓ_2 norm. The standard ℓ_1-minimization problem solves the following convex program:

$$\text{minimize} J(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1 \text{ subject to } \mathbf{x}' = \mathbf{\Omega}(\mathbf{D}\boldsymbol{\alpha} + \mathbf{m}) \tag{6}$$



(a) A given signal $\mathbf{x}$ (b) Fourier coefficients $\boldsymbol{\alpha}$

Fig. 6. The original signal $\mathbf{x}$ (a) and its frequency domain representation $\boldsymbol{\alpha}$ (b). (a) also depicts the location of 50 samples.

Eq. (6), known as *basis pursuit* (BP) [44], finds the vector with smallest ℓ_1 norm of vector $\boldsymbol{\alpha}$ defined as $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^{d} |\alpha_i|$. In ℓ_1-minimization literature, underdetermined problems are usually the norm rather than the exception, and N can largely exceed d'. Moreover, Eq. (6) is known to have a number of desirable attributes, such as its ability to return a sparse solution, and its numerical stability. As will become evident in the following sections, we will heavily rely on these advantages to achieve robust sparse feature extraction for both analysis and synthesis. As the results in [45] show, if a sufficiently sparse $\boldsymbol{\alpha}_0$ exists such that $\mathbf{x}' = \boldsymbol{\Omega}(\mathbf{D}\boldsymbol{\alpha}_0 + \mathbf{m})$, then Eq. (6) will find it. In the presence of noise, Eq. (6) becomes:

$$\text{minimize } \|\boldsymbol{\alpha}\|_1 \text{ subject to } \left\|\mathbf{x}' - \boldsymbol{\Omega}(\mathbf{D}\boldsymbol{\alpha} + \mathbf{m})\right\|_2 \leq \epsilon \tag{7}$$

for a given ϵ and is known as *basis pursuit denoising* (BPDN) [46] in the signal processing community and *Lasso* [47] in the statistical community. Let $\hat{\boldsymbol{\alpha}}_{\ell_1}$ denote the solution we obtain from either Eqs. (6) or (7), and the corresponding synthesis vector $\hat{\mathbf{x}}_{\ell_1} = \mathbf{D}\hat{\boldsymbol{\alpha}}_{\ell_1} + \mathbf{m}$. Compared to many greedy pursuit algorithms commonly used in ℓ_0-based dictionary learning approaches [48,49], the ℓ_1-based method enjoys significant improvement in terms of learning efficiency. Compared to many ℓ_2-based subspace learning methods, the ℓ_1-based method shows outstanding robustness towards noise and outliers.

3.2. Super-resolution from a signal processing and compressed sensing perspective

From a signal processing perspective, our aim is to reconstruct a signal from a number of samples much smaller than the Shannon–Nyquist criterion for signal recovery. Compressed sensing (CS) theory established how ℓ_1-minimization is able to recover the signal from partial sampling, by exploiting one fundamental assumption: that we have domain-specific knowledge of the signal being sampled. In our case, we know that the signal represents a face, and that signal has a sparse expansion in some basis.

To illustrate this fundamental concept, we briefly demonstrate it on a toy example. The concept is to sample significantly less than the Nyquist criterion and yet recover the signal. Let $\mathbf{x}$ be a signal synthesized by the addition of 4 complex sinusoids[2]. Fig. 6(a) depicts $\mathbf{x}$ while its frequency domain representation, $\boldsymbol{\alpha}$, is shown in Fig. 6(b).

We know that $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ is the Fourier coefficients and $\mathbf{D}$ is the Fourier basis dictionary.[3] $\mathbf{D}$ in this case is an orthonormal basis and finding $\boldsymbol{\alpha}$ for a completely sampled signal $\mathbf{x}$ reduces to a simple projection operation, given by Eq. (5). For the ℓ_2 formulation to work, Nyquist dictates that we need to sample at least twice the highest frequency. However, in our super-resolution scenario, we can only observe a fraction of the samples. Assume we observe only 50 out of the 1000 samples (depicted in Fig. 6(a) by the red circles), the ℓ_2-minimum norm solution fails to recover the original frequencies that originally generated the signal $\mathbf{x}$. Figs. 7(a) and 7(b) depict the degenerate analysis and synthesis steps respectively.

The main reason why ℓ_2-minimum norm failed is that $\boldsymbol{\Omega}\mathbf{D}$ is no longer orthogonal and there are many solutions that satisfy this equation. On the other hand, by recognizing that $\boldsymbol{\alpha}$ is sparse in the Fourier domain *and* forcing a sparse solution via ℓ_1-minimization, we can accurately recover the frequencies and reconstruct the signal using Eq. (6) as shown in Fig. 7(c) and 7(d). This remarkable results of completely reconstructing a signal with just 5% of the samples is at the core of our super-resolution technique. We build a multi-resolution face model where signals have an inherent sparse expansion and use ℓ_1-minimization to recover the entire original signal.

Formally, assume that a vectorized image $\mathbf{x}$ of size d needs to be recovered from d' measurements (such that $d' \ll d$). If we can also assume $\mathbf{x}$ has a sparse linear expansion in a basis $\boldsymbol{\Psi}$ such that $\mathbf{x} = \boldsymbol{\Psi}\mathbf{c}$ and that $\mathbf{c}$ is S-sparse. Assume only d' total measurements of $\mathbf{x}$ are observed, where every measurement i is obtained by the inner product of the i^{th} row $\boldsymbol{\phi}_i$ of a measurement matrix $\boldsymbol{\Phi}$ with the image, given by $\langle \boldsymbol{\phi}_i, \mathbf{x} \rangle$. The coherence of the two bases, given by $\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi}) = d \cdot \max_{i,k} \left|\langle \boldsymbol{\phi}_i, \boldsymbol{\psi}_k \rangle\right|$, where $\boldsymbol{\psi}_k$ is the kth column of $\boldsymbol{\Psi}$, was introduced in [50] to represent how "distant" the two basis are from each other. It was shown in [50] that if d' is greater than $\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi}) \cdot S \cdot \log(d)$, then ℓ_1-minimization will recover the image $\mathbf{x}$ with very high probability.

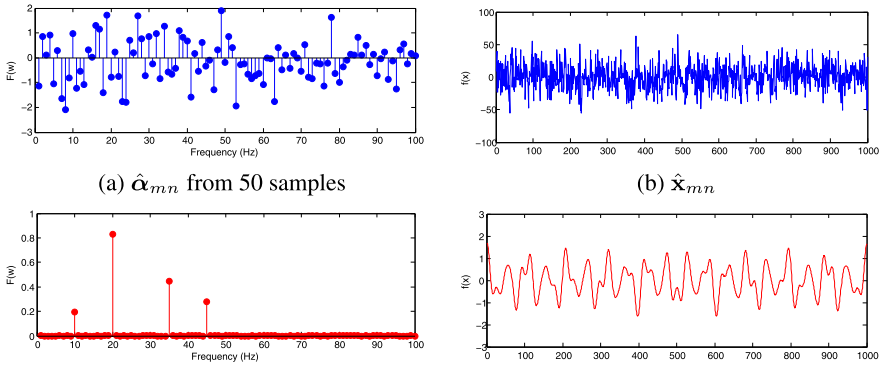
From this incoherence view of compressed sensing, we can justify why our derivation can recover the full multi-resolution pyramid. In our formulation, the sparse basis $\boldsymbol{\Psi}$ is the matrix of SVD dictionary atoms $\mathbf{D}$. The image $\mathbf{x}$ has a fundamentally sparse basis expansion in $\mathbf{D}$ as most of the energy is concentrated in the first few dictionary atoms. The original measurement basis $\boldsymbol{\Phi}$ is the standard canonical basis. In the presence of missing pixels, the measurement basis is represented by our indicator matrix $\boldsymbol{\Omega}$. The coherence between these two bases is minimal, and CS theory indicates that despite a significant proportion of missing pixels, we can still reconstruct the original image $\mathbf{x}$ using ℓ_1-minimization as we have shown. One could learn a purpose-specific dictionary that might generate a better reconstruction, but an SVD-type dictionary offers a sparse enough representation that ℓ_1-minimization can recover. As for the mechanics of the ℓ_1 solver, there are several classes of algorithms that seek to solve Eq. (7). The intrinsic details of specific solvers is beyond the scope of this paper, however, a solver based on the augmented Lagrangian method (ALM) [51] has so far consistently outperformed other solvers for our super-resolution application, and will be used for the remainder of this study.

3.3. Super-resolution induced by sparse representation

The high-quality reconstructions achieved by ℓ_1-minimization on the occluded faces are significant because, for super-resolution, the reconstruction task is even more challenging. Comparatively, a bigger proportion of the pixels is missing, and we need a method that is able to fit a signal representation vector from very few observed pixels and reconstruct the entire pyramid. In this section,

[2] assume $\mathbf{x}$ is of length N=1000 and the generating frequencies are (10 Hz, 20 Hz, 35 Hz, 45 Hz) with respective amplitudes of (0.3 0.9 0.5 0.4).

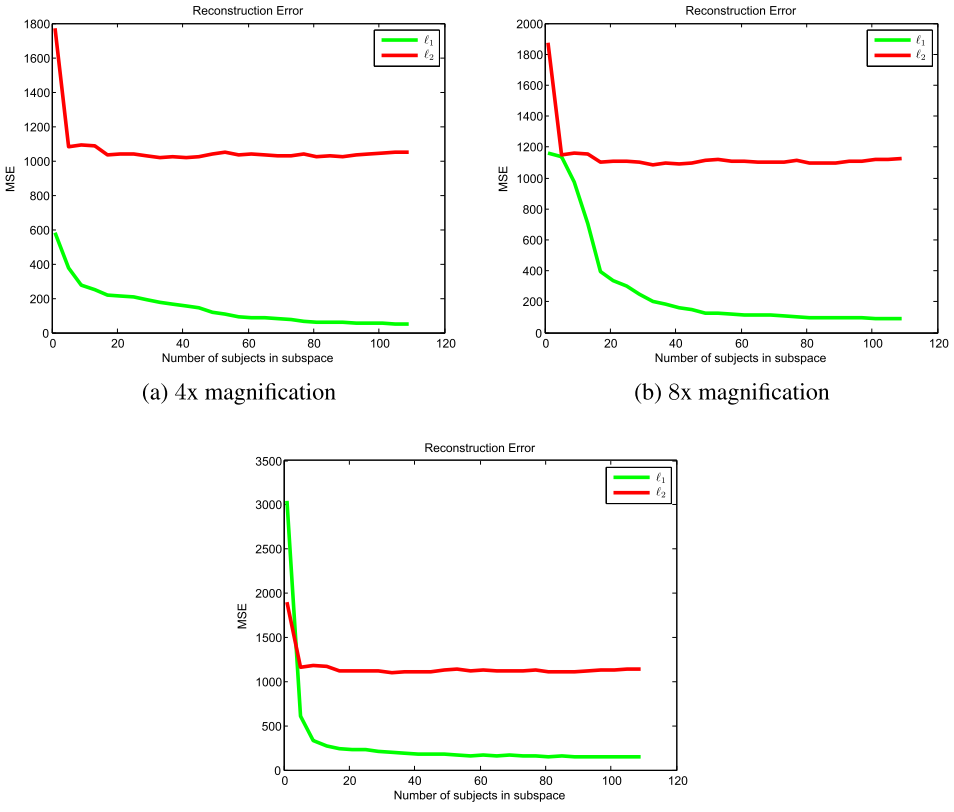
[3] $\mathbf{D}_{p,q} = \frac{1}{\sqrt{N}} e^{-2\pi jpq/N}$.



(a) $\hat{\boldsymbol{\alpha}}_{mn}$ from 50 samples (b) $\hat{\mathbf{x}}_{mn}$

(c) $\hat{\boldsymbol{\alpha}}_{\ell_1}$ from 50 samples (d) $\hat{\mathbf{x}}_{\ell_1}$

Fig. 7. Analysis and synthesis results using ℓ_2 and ℓ_1 minimization respectively. The former fails with so few samples while the latter completely recovers the original signal.



(a) 4x magnification (b) 8x magnification

(c) 16x magnification

Fig. 8. Mean square error between the reconstructed high-resolution face belonging to an unseen subject and the true high-resolution face produced by the ℓ_1 and ℓ_2 methods using an increasing number of subjects in the training subspace. (a), (b) and (c) show the MSE plots at different magnification factors.

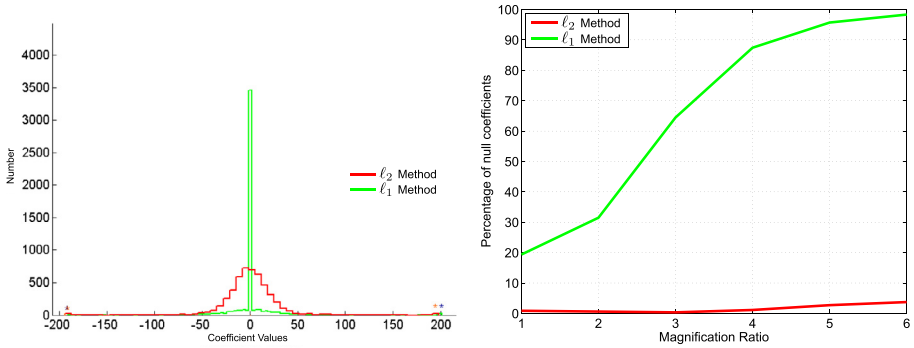
we empirically show how sparsity is essential in achieving this aim.

3.3.1. Robustness analysis

The main advantage of the sparse feature extraction method over the ℓ_2 methods lies in its ability to find *and* maintain a "good" solution, regardless of the size of the training data. To illustrate this concept, we build a multi-resolution subspace consisting of an increasing number of training subjects, all different from the test subject. We then solve for the coefficients using the ℓ_1 and ℓ_2 methods starting from different levels of the pyramid, inducing different magnification factors. Fig. 8 compares the mean squared error (MSE) between the original and reconstructed images as the number of training subjects increase. What we observe is that ℓ_1 achieves and maintains a consistently lower MSE regardless of the number of training subjects in the subspace.

3.3.2. Sparsity analysis

The key observation to make is that ℓ_1 achieves a smoother reconstruction because, when the number of missing dimensions is high, it aggressively forces the coefficient vector to be sparse. To quantify this observation, Fig. 9(b) depicts the distributions of coefficient values for a given magnification ratio. The coefficient value histograms of ℓ_2 get wider and wider as the level of occlusion increases, which means that it is struggling to fit a coefficient vector that best explains the observed data. Fig. 9(b) plots the



(a) (b)

Fig. 9. Sparsity Analysis. Figure (a) plots the frequency of coefficient values given a magnification factor for the different methods. Figure (b) plots the sparsity ratio or percentage of null coefficients as a function of the following magnification factors: 1x, 2x, 3x, 4x, 5x, and 6x.

percentage of null coefficients for the two methods as the magnification factor increase. This result suggests that ℓ_1 abstains from assigning weights arbitrarily when the fitting becomes challenging as a result of missing dimensions. This in return enables it to maintain a natural looking reconstruction, because higher-order dictionary atoms that represent mostly noise are not assigned a heavier weight as the system becomes increasingly underdetermined.

For the remainder of the paper, we will drop the ℓ_2 methods and report results for the ℓ_1 only, which are consistently better.

3.4. Discussion

Our low-resolution parent-structure feature extraction step is similar to the one described in [3], but unlike in [3], our method is global and relies on the ℓ_1-minimization to avoid degenerating towards a mean face. Unlike [15] where the coefficient vector $\mathbf{c}$ is extracted by projecting on low-resolution eigenfaces and reconstructing using high-resolution eigenfaces (and which does not provide any theory as to what ties the two subspaces together), our subspace learns the coupled high-and-low frequency from the training Gaussian pyramids (or equivalent features) simultaneously. Consequently, our dictionary atoms are multi-resolution, hence no artificial constraints need to be introduced to guarantee a smooth face-like reconstruction.

More importantly, the simplicity of our method keeps it modular and allows for further improvements. For instance, it could replace the global reconstruction part of any two-step approach (*e.g.*, [14,16]) which first compute a rough global reconstruction and further enhance it with a slower and more specialized local patch-based algorithm. Similarly, the Gaussian pyramid-based subspace could be augmented with any other features such as a Laplacian pyramid or gradient-based features.

Readers are encouraged to check out an earlier work by Abiantun et al. [35], which is closely related to this paper. In [35], the authors have proposed a missing data recovery paradigm for solving face pose correction problem. In our work, we are tackling the super-resolution problem also from a missing data recovery point of view. Although the underlying assumption is shared, these two works take on very different computer vision problems (pose correction vs. super-resolution), and in order to be better suited for the super-resolution task at hand, we have incorporated components and solutions that are unique to the super-resolution problem such as image pyramid representation in shape-free domain for tackling various low-resolution cases, as well as an improved 3DGEM 3D modeling technique for handling low-resolution face landmarking, *etc.*

4. Evaluation

4.1. Dataset

The training face images to be used in our experiments are sourced from the following databases: (1) the CMU Multi-PIE (MPIE) [39] database Sessions 2 and 3, (2) the Face Recognition Grand Challenge (FRGC) ver2 [52] database, and (3) the FERET [53] database.

The MPIE database [39] contains more than 750,000 images of 337 people recorded in up to four sessions over the span of five months. Subjects were imaged under 15 view points and 19 illumination conditions (with 1 neutral as well) while displaying a range of facial expressions. In our experiment, we will only use frontal images with neutral illumination and expression.

The FRGC ver2 database [52] has three components: first, the generic training set is typically used in the training phase to extract features. It contains both controlled and uncontrolled images of 222 subjects and a total of 12,776 images. Second, the target set represents the people that we want to find. It has 466 different subjects and a total of 16,028 images. Last, the probe set represents the unknown images that we need to match against the target set. It contains the same 466 subjects as in target set, with half as many images for each person as in the target set, bringing the total number of probe images to 8014. All the probe subjects that we are trying to identify are present in the target set. Images from the three sets are mutually exclusive.

The FERET database [53] was collected in 15 sessions. The database contains 1564 sets of images for a total of 14,126 images that includes 1199 individuals.

We train our SSR2 model using a mixture of face images from the aforementioned databases which amount to over 40,000 frontal face images. We call this mixture database *SSR2-DB*. The testing face images are from the MPIE session 1, with non-overlapping subjects distinct from the training stage.

4.2. Protocol

We evaluate the ℓ_1 reconstruction for global face hallucination by downsizing the original MPIE session 1 images in the shape-free domain. The original shape-free images provide an interocular distance (IOD) of 100 pixels, and when we reconstruct a lower resolution image, we recover that original size. We evaluate the quality

Table 1
Summary of the average PSNR (dB) and average SSIM for different super-resolution techniques.

		Bi-cubic [7]	B-spline [8]	Lanczos3 [61]	Yang et al. [16]	Kim et al. [18]	SRCNN [30]	SelfEx [11]	SRGAN [34]	Ours
IOD		Average PSNR (dB)								
25 pix	4x	24.21	24.23	24.32	24.53	24.37	24.48	24.51	24.82	**32.80**
12.5 pix	8x	20.59	21.04	20.66	20.77	20.57	20.71	20.62	20.78	**28.97**
6.25 pix	16x	8.47	19.14	8.46	17.84	17.66	17.90	17.77	17.95	**26.33**
		Average Structural Similarity (SSIM)								
25 pix	4x	0.6872	0.6889	0.6920	0.7003	0.6958	0.6976	0.6986	0.7113	**0.9184**
12.5 pix	8x	0.5480	0.5668	0.5530	0.5576	0.5472	0.5541	0.5503	0.5560	**0.8355**
6.25 pix	16x	0.1075	0.4847	0.1081	0.4245	0.4158	0.4273	0.4231	0.4317	**0.7584**

of our reconstruction for different magnification factors (different k-levels of the Gaussian pyramid).

4.3. Metrics

The peak signal-to-noise ratio (PSNR) is a common objective image reconstruction quality metric typically used in denoising and image/video compression applications [54–59]. As the MSE approaches zero, the PSNR goes to infinity. In lossy image/video compression, the typical PSNR values range from 30 to 50 dB, and anything below 20 dB is deemed unacceptable [60]. As is also common, since the PSNR is a logarithmic scale, the average PSNR reported was computed by first measuring the average MSE and then converting the average MSE to PSNR rather than averaging PSNR values directly. The structural similarity (SSIM) is used here as a secondary metric for completeness.

4.4. Benchmarks and super-resolution experimental results

We benchmark results against the best interpolation techniques, such as bicubic polynomial interpolation [7] and Lanczos resampling [61] which uses a Lanczos kernel (a windowed sinc function) to smoothly interpolate the value between samples. This latter method is usually employed by most commercial photo displaying software. We also benchmark against cubic B-spline interpolation [8] which marginally outperforms traditional bicubic polynomial interpolation. Table 1 summarizes the average PSNR and average SSIM obtained for 249 reconstructions using the different methods.

We also benchmark our method for super-resolution against the method in Yang et al. [16] (an implementation of the approach is available [62]) since theoretically this approach is the closest to ours. The differences are that our approach is global, and our dictionary is given by Gaussian-pyramid face model, while theirs is twofold, one global based on NMF to generate a smooth intermediate face, and then an expensive local patch-based approach to infer high-frequency information. The intermediate global face provides patches whose sparse representation (from a low-resolution dictionary) is used to generate a high-resolution patch (using the high-resolution dictionary, and both dictionaries are learned jointly).[4] Since their method is local patch-based, their storage requirement is much lower than our method, since they only need to keep the two compact dictionaries. On the other hand, their method is much slower because of the number of overlapping patches it needs to process, while our method extracts the sparse representation once from the entire low-resolution face. Even though their method marginally beats interpolation methods at 4x magnification, it breaks down for lower input resolutions and returns a very blurry reconstruction.[5]

Next, we benchmark against the method of Kim et al. [18]. An open-source implementation by the author is available [63]. However, we could not retrain their algorithm on face images only so we used it as is. It is important to note that their magnification factor was limited to 4x, so for higher magnification rates we ran the algorithm successively twice to achieve higher magnifications.

Moreover, we benchmark against recent deep learning based SRCNN [30,31] method and SRGAN [34] method, as well as the self exemplar (SelfEx) based [11] super-resolution method. SRCNN and SelfEx methods have implementations by the authors available. SRGAN has third-party implementations available [64,65]. For the SRCNN, the scaling factor available is 2x, 3x, and 4x, so for our 8x SR experiments, we perform 2x followed by 4x, and for 16x SR experiments, we perform 4x followed by 4x. The SRCNN is a patch-based deep learning algorithm, which is trained on 91 natural images with roughly 24,800 sub-images. In this work, we perform fine-tuning using face image patches on top of the original SRCNN which is only trained on natural images. We use around 25,000 face image patches extracted from the SSR2-DB in the shape-free domain, which is about the same number of patches used in the original SRCNN training. For self exemplar based method, the highest scaling factor is 8x, so for our 16x magnification experiments, we perform 2x followed by 8x. For the SRGAN, the original work only deals with 4x scaling factor, and both 4x and 8x are dealt with in [65]. So for our 4x and 8x SR experiments, we use the options provided as is, and for 16x SR experiments, we perform 4x followed by 4x. The original SRGAN is trained on 350,000 images from ImageNet database [66], and to make it better suited for our tasks, we fine-tune the SRGAN using the entire SSR2-DB database with over 40,000 frontal face images in the shape-free domain.

The results tabulated in Table 1 represent the average PSNR in dB and average SSIM produced by different magnification rates (starting from different interocular distances) and for different techniques. Visual results are shown in Fig. 10. As can be seen, the proposed SSR2 method significantly outperforms the other competing algorithms both qualitatively and quantitatively. It is worth noting that although we have fine-tuned the SRCNN using images in the face domain, the algorithm is still not able to super-resolve sharp images, especially when the zooming factor is large. For SRGAN, in lower magnification cases, faces are successfully super-resolved, but the method is still limiting when dealing with larger magnification factors. For self exemplar based method, it seems that face images do not exhibit salient self exemplars under affine

[4] We retrained the method of [16] using the software made publicly available by the first author using different crops and different number of faces, and different dictionary sizes. The results we show use the following parameters: $\lambda = 0.1$, 100,000 patches, dictionary size = 1024, upscaling factor 2, patch size = 5, overlap = 4, 1000 shape-free tightly-cropped and registered training faces. The results reported were obtained without the use of their backprojection global method which seemed to hurt the MSE of the reconstructed high-resolution face image.

[5] We tried different product of upscale ratios, for example to achieve 8x, upscale with a factor of 2 three times, or with a factor of 4 followed by a factor of 2.

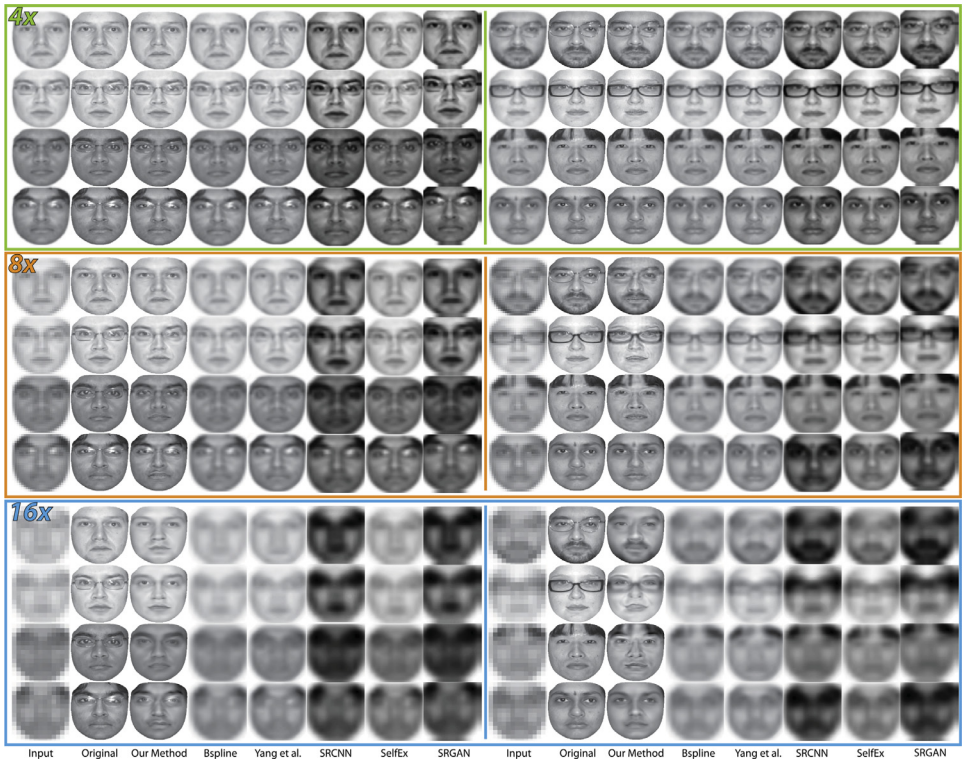


Fig. 10. Super-resolution results on test faces. Each set of 8 columns depicts (L to R) input image, original face, our reconstruction, cubic B-spline [8] reconstruction, the method of Yang et al. [16], SRCNN [30,31], self exemplar based (SelfEx) method [11], and SRGAN [34] method. The top four rows show 4x magnification results (25 pixels IOD), middle four rows show 8x magnification results (12.5 pixels IOD), and the last four rows show 16x magnification results (6.75 pixels IOD).

transformations, which is quite different from images of man-made objects and buildings in an urban environment [11].

4.5. Sensitivity to image noise

So far we have assumed that the input signal is *noise-free*. However, this is rarely the case when super-resolution is needed, as low-resolution faces usually originate from poor quality surveillance cameras or video footage. In these cases, the signal will most likely be contaminated with noise. One could first denoise the image and then apply super-resolution. However, we can demonstrate that our approach can intrinsically handle noise without an explicit denoising step. Let's assume that our signal $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha} + n$ is corrupted with additive white Gaussian noise n (AWGN is white noise with a constant spectral density and a Gaussian distribution of amplitude). Let the noise be normally distributed with zero mean and variance σ^2. Eq. (7), after dropping the indicator matrix $\boldsymbol{\Omega}$ for notational simplicity, can be reformulated using Lagrange multipliers.

$$\arg\min_{\boldsymbol{\alpha}} \|\mathbf{D}\boldsymbol{\alpha} - \mathbf{x}_{\boldsymbol{\alpha}}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{8}$$

This in return corresponds to the traditional Bayesian formulation that seeks to maximize the posterior probability:

$$P(\boldsymbol{\alpha}|\mathbf{x}) = \frac{P(\mathbf{x}|\boldsymbol{\alpha})P(\boldsymbol{\alpha})}{P(\mathbf{x})} \tag{9}$$

which corresponds to solving the following MAP problem:

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} P(\boldsymbol{\alpha})P(\mathbf{x}|\boldsymbol{\alpha}) \tag{10}$$

where the prior $P(\boldsymbol{\alpha})$ is assumed to be Laplacian with location parameter 0 and scale parameter b given by:

$$P(\boldsymbol{\alpha}) = \frac{1}{2b}\exp\left(-\frac{\|\boldsymbol{\alpha}\|_1}{b}\right) \tag{11}$$

Since $n \sim \mathcal{N}(0, \sigma^2)$ the likelihood $P(\mathbf{x}|\boldsymbol{\alpha})$ is also normally distributed with mean $\mathbf{D}\boldsymbol{\alpha}$ and variance σ^2:

$$P(\mathbf{x}|\boldsymbol{\alpha}) = \frac{1}{2\sigma^2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{D}\boldsymbol{\alpha} - \mathbf{x}\|_2^2\right) \tag{12}$$

Combining Eqs. (12) and (11) back in Eq. (9) translates into maximizing the sum of the exponents, or minimizing the negative of the sum of the exponents:

$$\arg\min_{\boldsymbol{\alpha}} \frac{1}{2\sigma^2}\|\mathbf{D}\boldsymbol{\alpha} - \mathbf{x}\|_2^2 + \frac{\|\boldsymbol{\alpha}\|_1}{b} \tag{13}$$

Eq. (13) and (8) are identical for $\lambda = \sigma^2/b$. This shows that our technique is inherently a Bayesian approach that is designed to handle noise by modeling the prior of the coefficient vector $\boldsymbol{\alpha}$, assuming white Gaussian noise, and looking for the solution vector that is optimal is the MAP-sense. λ controls the sparsity of the solution, so the sparser the model, the bigger σ^2 can be (assuming b to be constant) which means the more noise in the data our model assumes. Fig. 11 depicts noisy image super-resolution reconstruction. The images have been corrupted by the AWGN channel with variance increasing by a factor of 10 in each case ($\sigma^2 = 0.0001$, 0.001 and 0.001 respectively). The parameter ν that controls the sparsity of the ℓ_1 solver has to be increased accordingly. We select the optimal ν by finding the optimal value with respect to reconstruction error on a validation set. As predicted by our above calculation, those optimal values are greater by a factor of 10 each time, confirming our model.

This last analysis highlights the similarities between our method and standard Bayesian super-resolution approaches [3,67–70]. The most significant contrast is that their methods operate in pixel space (or a related space, such as Laplacian or gradient features), while we use an SVD dictionary of a shape-free Gaussian pyramid representation. Moreover, our smoothness prior is an ℓ_1-based measure which prevents the reconstruction from

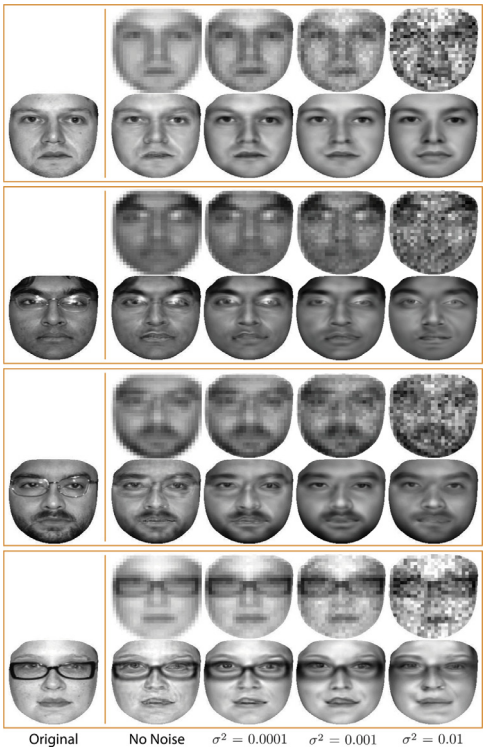


Fig. 11. Noise Tolerance. Reconstruction with 8x magnification, starting from an interocular distance of 12.5 pixels. The low-resolution images were corrupted with AWGN of increasing variance.

degenerating towards the average face, which is often the case in traditional Bayesian super-resolution techniques.

4.6. Sensitivity to landmarking noise

Another common source of signal noise is introduced by the reliance of the proposed technique on landmarking errors [71,72]. Our experiments have shown that even among human annotated images, the expected amount of variation in the fiducial point locations is around −3 dB. To evaluate the impact of such noise on the system, we perturbed the manually labeled landmarks with artificial AWGN of increasing level and computed the impact on the reconstructed faces as PSNR. The result of this experiment is shown in Fig. 12.

Naturally, the introduction of landmarking noise results in an expected reduction in the PSNR of the resulting super-resolution, regardless of the scale of the resolution. Crucially, we observe a graceful degradation in the system performance with increasing noise, and the decrease is minimal up to −5 dB which represents a significant perturbation in the landmarks that is perceptually worse than a bad human input. As the amount of noise increases, we note that lower resolutions are affected less by the introduced noise, perhaps due to the limited impact of this landmarking error on the already poor textural quality of the image. We find that at approximately −11 dB of noise (which is unrealistic in most real-world applications) the impact on the resulting texture perturbation in higher resolutions (4x, 8x) causes the PSNR to drop significantly. However, the low-resolution (16x) result is still stable.

4.7. Face recognition with super-resolution

The MSE, PSNR, and SSIM metrics quantify the faithfulness of the reconstruction to the original face. Our next experiment measures how much of the discriminating information of the original

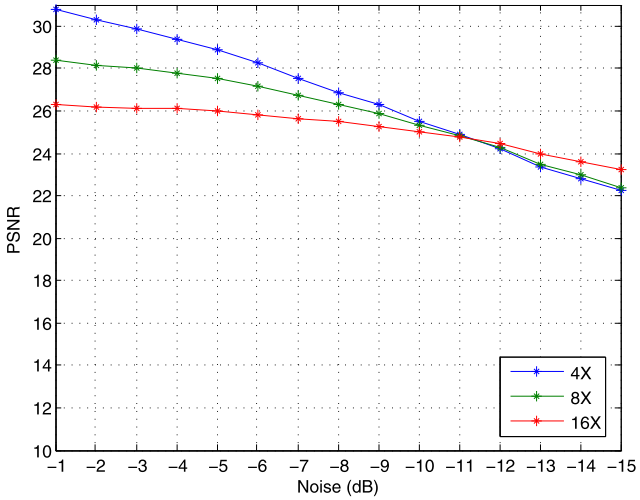


Fig. 12. Sensitivity of the technique to artificially induced landmarking errors, for the magnification factors of 4x, 8x and 16x.

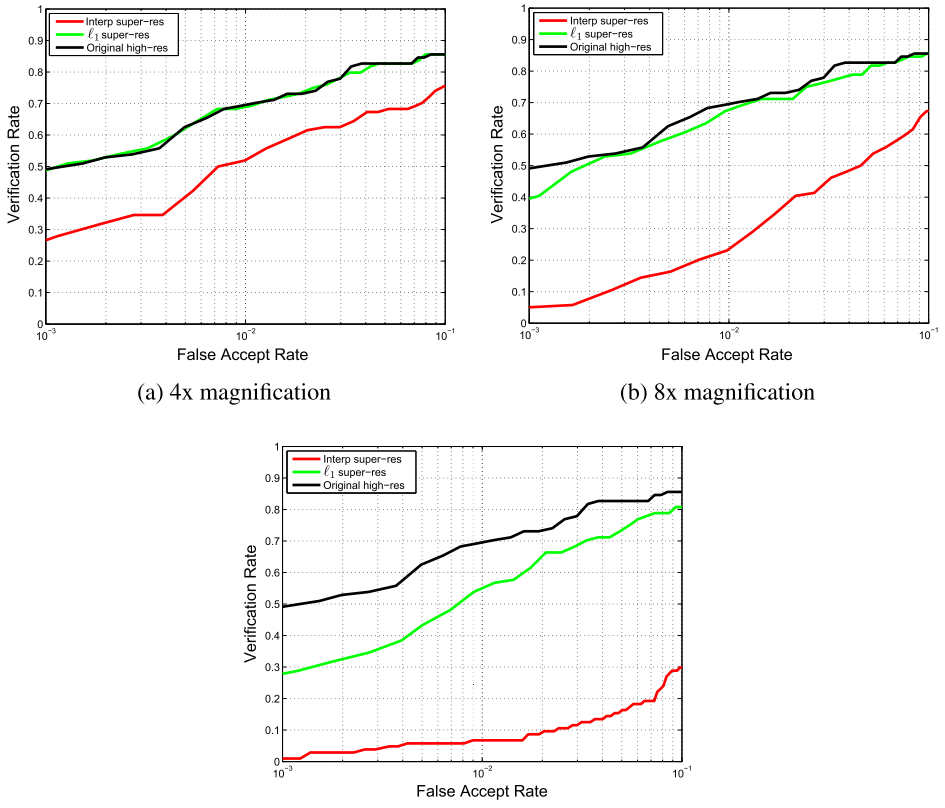
face the reconstruction retains. For that, we set up a simple face matching experiment, and measure the verification rate at different magnification levels, summarized by the receiver operating characteristic (ROC) curves in Fig. 13. The gallery set consists of the original high-resolution MPIE session 1 shape-free faces, while the query set contains 341 subjects with 104 of them seen in the gallery set (same subjects, but different images as they originate from MPIE session 2) while the rest of them are unseen in the gallery set (a mix of MPIE session 2, session 3, and FERET [53] faces). The input query face images have been downsized to provide 25, 12.5 and 6.25 pixels between the eyes and then reconstructed to the original 100 pixels between the eyes, to provide the magnification ratios of 4, 8 and 16 respectively. Normalized cosine distance [73–77] is the metric used for the simple matcher. Fig. 13 shows that ℓ_1 sparse representation recovery yields hallucinated faces with 4x magnification that offer the same discrimination between faces as the original high-resolution images. Naturally the ROCs drop for high magnification ratios, but the relative drop for our method is significantly less than for naive interpolation using cubic B-spline [8].

4.8. Occlusion robust super-resolution

Degradations such as low-resolution and facial occlusions usually come in a bundle. It is desirable for a super-resolution algorithm to be able to handle occlusion removal as well. Due to the formulation of the proposed algorithm, where the super-resolution task is cast as a missing data challenge, the same algorithm, in nature, can simultaneously deal with occlusion removal as well. The only thing required by the algorithm is a user-specified binary occlusion mask which translates to a binary row selection matrix $\mathbf{\Omega}_{\text{occlusion}}$. By taking the Hadamard product $\circ$ between $\mathbf{\Omega}_{\text{occlusion}}$ and the original super-resolution binary row selection matrix $\mathbf{\Omega}$ as used in Eq. (2), we obtain a new binary row selection matrix $\mathbf{\Omega}_{\text{multitask}}$ that captures both the missing data from the super-resolution task as well as the occlusion removal task

$$\mathbf{\Omega}_{\text{multitask}} = \mathbf{\Omega}_{\text{occlusion}} \circ \mathbf{\Omega} \tag{14}$$

Since other competing algorithms cannot deal with super-resolution and occlusion removal simultaneously, we only some qualitative results in Fig. 14. The multitask results are obtained from two artificially added facial occlusions on the same face image, at the same extremely low resolution (16x case), as can be seen in Fig. 14(a) and 14(d). User-specified occlusion masks are



(a) 4x magnification

(b) 8x magnification

(c) 16x magnification

Fig. 13. Effect of super resolution on cross-session matching. The gallery images are 249 full-resolution face images from MPIE session 1. The query images are 341 reconstructed super-resolution images from MPIE sessions 2, 3 and FERET. 104 of test subjects are seen in the gallery set. Our face hallucination method handles the drop in resolution much more gracefully than bicubic interpolation. With an input resolution of 25 pixels between the eyes, our super-resolution technique fares as well as the original high-resolution in this face verification experiment. Here the interpolation method is cubic B-spline [8].

provided to the algorithm and are visualized in Fig. 14(b) and 14(e). The simultaneous super-resolution and occlusion removal recoveries are shown in Fig. 14(c) and 14(f). When the occlusion is relatively small (Fig. 14(a)), the recovered face shows pretty high visual fidelity. When the occlusion is relatively large (Fig. 14(d)), the proposed algorithm is still able to recover a visually pleasing facial image, with details well preserved especially in the eye region since it is the only region visible in the occluded low-resolution image.

4.9. Discussion

Efficiency: As we have discussed in Section 3 and depicted in Fig. 2, the potential computation bottleneck of the proposed SSR2 method during testing (*e.g.*, deployed for real-world application) lies in (1) getting the shape-free representation and (2) getting the sparse coefficient vector through pursuit algorithms such as BP and BPDN that are based on ℓ_1 minimization.

As we have discussed in Section 3.1.1, we rely on an efficient 3D modeling technique called 3DGEM [37] for obtaining the shape-free representations of any given face. The 3DGEM method itself is real-time, and it only requires CPU computation. A recent follow-up work [78] of 3DGEM can achieve much faster 3D face modeling (faster than real-time) using GPU. This method uses 3D spatial transformer networks with thin plate spline (TPS) warping to generate both a 3D model of the face and accurate 2D landmarks across large pose variation.

As we have discussed in Section 3.2, we use the augmented Lagrangian method (ALM) [51] for solving the ℓ_1 minimization problem in Eq. (7). When evaluated on a 3.40 GHz Intel Core i7-4770 CPU, the ALM takes around 0.1 s. for each sparse signal recovery.

Other components in the SSR2 method are mainly linear algebra operations which are not computationally expensive. Therefore, the SSR2 approach as a whole is a fairly efficient face super-resolution algorithm when deployed for real-world applications.

The role of deep learning: Here, we want to briefly discuss whether deep learning techniques can be combined with the proposed method for obtaining even better performance. Shallow methods such as the one presented in this work can effectively handle constrained problems without too much variation, although we have demonstrated some level of robustness towards noise and misalignment. On the other hand, deep learning methods are very well suited for handling unknown variations such as pose, illumination, and expression *etc.*, as well as unknown degradations such as noise and occlusion. One way to combine deep learning with the proposed method is to apply deep learning-based normalization technique as a pre-processing step to normalize and pose, illumination, expression, *etc.*, and remove various degradations, so that the subsequent super-resolution on faces with extreme low resolutions can perform well, especially for less constrained real-world scenarios.

5. Application to a real-world scenario

The face recognition results of the previous section use down-sampled versions of high-quality images to simulate lack of resolution. In this section, we show that our method generalizes to real-world grainy and low-quality surveillance images, where super-resolution will most likely be applied and can have its biggest impact. In the wake of the tragic terrorist bombing that hit the 2013 edition of the Boston Marathon, we got to test our super-resolution technique on real-world surveillance footage that

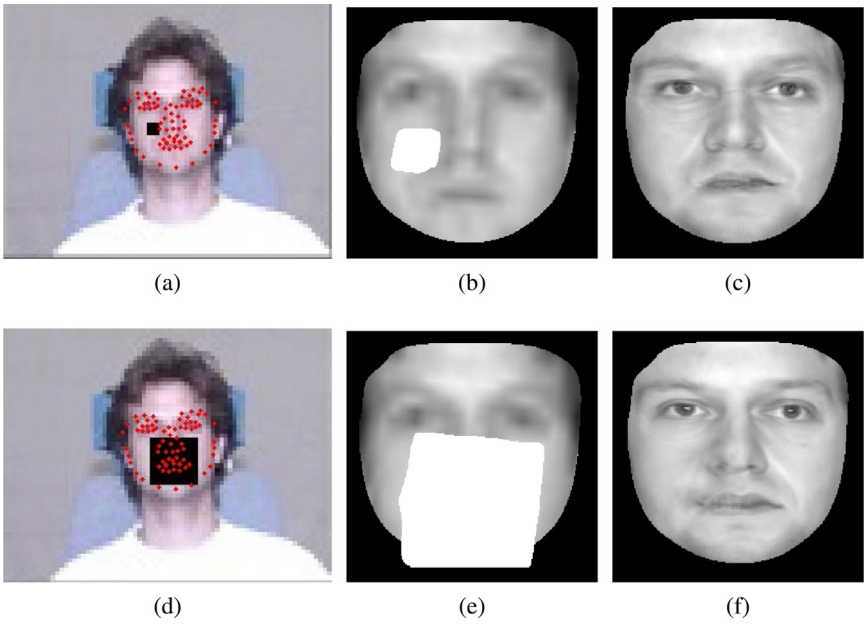


Fig. 14. (a,d) Low-resolution face with small and large occlusion respectively; (b,e) user-specified occlusion masks; (c,f) super-resolution recovery (16x) with occlusion removal capability.

Fig. 15. Left: Immediate aftermath of the first blast (within 10 s)[84]. Right: The blasts occurred close to the finish line (yellow) along the marathon course (dark blue), with the first blast being closer to the finish [84] (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

Fig. 16. FBI wanted posters for two suspects released on April 18, 2013.

was made available by the FBI in their hunt for the suspects [1]. The CMU super-resolution enhanced image of the bombing suspect was sent over to the FBI at late night of *April 18, 2013*, as the event was unfolding. It was before the authorities have positively identified the Dzokhar brothers. Our involvement in the identification of the Boston Marathon bombing suspects was later reported by Miles O'Brien in a documentary by PBS NOVA: "Manhunt: Boston Bombers" [79–81] with excerpts in Fig. 17.

During the annual Boston Marathon on *April 15, 2013*, two pressure cooker bombs exploded at *2:49 p.m. EDT* near the finish line, killing 3 spectators and injuring 264 others. The immediate scene after the first blast and the locations of two blasts are shown in Fig. 15. At *April 18, 2013, 5:00 p.m. EDT*, the FBI released surveillance video[6] and photographs of two suspects in the bombing case, as shown in Fig. 18 [1], along with the wanted posters shown in Fig. 16(a) and 16(b), according to FBI updates [83]. In Fig. 18, the only usable image is Image #11 (officially referred to as "hand-by-ear-circled" by the FBI), but still, it is of poor quality with low resolution, off-angle pose, and occlusions. At *April 19, 2013, 8:20 a.m. EDT*, suspect No.2 was positively identified as shown in Fig. 1(d).

[6] The extremely low frame rate of the surveillance camera footage has rendered gait-based recognition method [82] useless.

There are more than 15 h gap between the initial photo release from the FBI till the final positive identification of suspect No.2, and during which one MIT police officer was shot dead by the suspect, followed by a carjacking. Should the suspects be sooner identified by facial recognition technology, these further tragedy might be avoided. It is the suspects' aunt who identified the brothers, not automated facial recognition systems. The current facial recognition system in place cannot handle unconstrained facial recognition scenario well, such as low resolution, off-angle pose, occlusions, *etc*. Given images of challenging quality like the ones shown in Fig. 18, the technology simply cannot return a high-confidence match due to the non-ideal quality of the query image, even though the photos of two suspects are all in government database in controlled enroll environment and available on social media.

5.1. Our single-image super-resolution enhancement

As seen in Fig. 18, all of the surveillance images released that day except two are useless for face recognition as a very small portion of the face is visible. Of the two remaining images, one

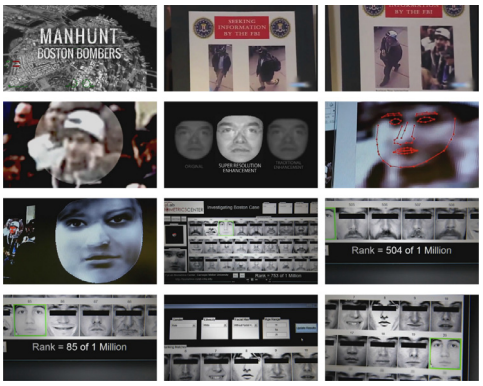


Fig. 17. Excerpts from the documentary by PBS NOVA: "Manhunt: Boston Bombers" reported by Miles O'Brien.

Fig. 18. FBI release on Thursday, April 18, 2013 showing all images that were obtained by the FBI from store-owned and public surveillance cameras.

was a full profile, so we used the last remaining image which depicted the suspect in a near-frontal viewpoint.

However, this original face in Fig. 18 (Image #11), as released by the FBI, presents approximately the equivalent of 7 *usable* pixels between the eyes. As can be seen, this particular degradation is more of a super-resolution recovery problem, rather than an occlusion removal problem such as recovering the full face image from a masked subject [49,85–87]. Although the proposed SSR2 method is able to handle both degradation s simultaneously as shown in Section 4.8, we did not provide occlusion mask in our effort to recover the Boston Marathon bomber face image for the reason discussed above. Using this face as input, we can hallucinate the high-resolution image shown in Fig. 1(c). This was the enhanced image sent over to the authorities on the night of *April 18, 2013* as the event was unfolding. Some manual preprocessing of the image was required to generate this image, especially for registration to properly align the image prior to enhancing the resolution. We also compare and match against a high-resolution mugshot-style face image of the suspect after he was identified and apprehended.

The hallucinated face in Fig. 1(c) bears a subtle but certain resemblance to the positively identified high-resolution face. To measure this similarity, we set up the following experiment: the positively identified photo of suspect No.2 (Fig. 1(d)) is included in the gallery set, along with one million mugshot images obtained from the Pinellas County Sheriff's Office (PCSO). Our enhanced image seen in Fig. 1(c) is the query image. This experiment aims to

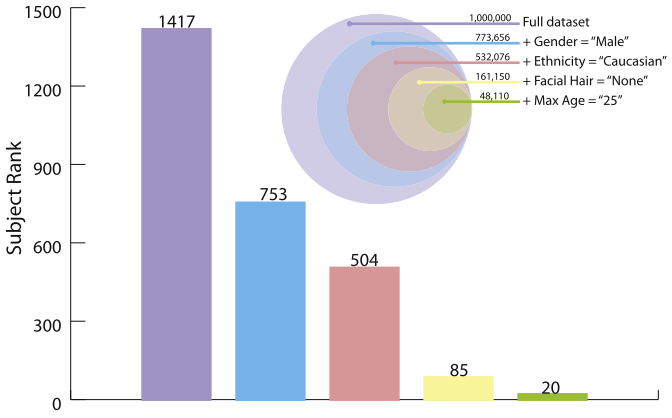


Fig. 19. Retrieval results when incrementally considering soft-biometrics attributes such as gender, ethnicity, facial hair, and age. Both the ranking and the corresponding pool sizes are shown. With all factors considered, we can locate the subject at rank 20.

simulate the automated facial matching procedure where only the unconstrained facial image of suspect No.2 (Fig. 18 Image #11) is available to the law enforcement. The purpose is to show how our enhancement and matching algorithm could have helped law enforcement to narrow down the search pool by retrieving the suspect from a fairly large pool with high retrieval ranking. As a comparison, Boston's city population in 2012 was 636,479 [88], much smaller than this gallery dataset.

The developed matching algorithm used is based on our KCFA algorithm [89] which ranked #1 in NIST's Face Recognition Grand Challenge (FRGC) [52,90] and trained on 12,776 face images of only 222 subjects from the generic training set of the FRGC database [91–94]. None of the one million gallery mugshot faces are included in this training set. To further mimic real-world police investigation scenarios, we narrow down the search space by considering available soft-biometric information [95], such as gender, ethnicity, facial hair, age, *etc*. All this information was available to the FBI from witness testimony and other images of the scene. Fig. 19 depicts our initial results and the improvement in identification results due to taking into account soft-biometric information. When disregarding matches that belong to the wrong ethnicity, gender, facial hair presence and age bracket, the enhanced image of the suspect was a *top 20 match*.

6. Conclusion

In this paper, we have outlined a novel super-resolution technique that we have shown to be effective in extreme low resolution imagery, able to achieve 16x resolution recovery of faces, which is significantly more than what has been demonstrated in the literature thus far. Our technique is based on reformulating the super-resolution problem as a missing data challenge, and we borrow ideas from compressed sensing to solve it. The reason we can sample significantly below Nyquist frequency and still recover the original signal stems from domain-specific knowledge of the low resolution signal. By employing an appropriate basis where our signal can be sparsely represented, we can recover this feature representation robustly and consistently using sparse recovery optimization techniques such as ℓ_1-minimization. We have benchmarked our method against ground-truth data and showed its effectiveness in reconstruction, and we have also demonstrated that this sparse feature recovered representation holds well in face recognition experiments. Moreover, the proposed super-resolution algorithm is, in nature, capable of simultaneous facial occlusion removal, a desirable property that other super-resolution algorithms do not possess. We have analyzed it on a real-world scenario, the

Boston Marathon bombing incident, and found that it could have been useful in identifying one of the suspects out of a database of a million mugshots. In this era where deep learning is the de facto panacea to all computer vision problems, we demonstrate in this paper that traditional method involving subspace modeling, sparse representation, and shape-free face pyramid can perform advantageously compared to more convoluted methods for face super-resolution problems with extreme low resolutions. In the future, we plan to extend this technique by incorporating tolerance to pose, occlusion and other degradations observed in real-world challenges that law enforcement can face.

Acknowledgments

This work was partially funded by CMU CyLab, and by the National Institute of Justice Grant no. 2013-IJ-CX-K005.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2019.01.032.

References

[1] FBI Photo Release on April 18, 2013. http://web.archive.org/web/20130418212959/http://www.fbi.gov/news/updates-on-investigation-into-multiple-explosions-in-boston/photos.
[2] J. Klontz, A. Jain, A case study of automated face recognition: the boston marathon bombings suspects, IEEE Comput. Mag. 46 (11) (2013) 91–94.
[3] S. Baker, T. Kanade, Hallucinating Faces, Technical Report, CMU-RI-TR-99-32. Robotics Institute, Pittsburgh, PA, 1999.
[4] R. Abiantun, M. Savvides, B. V. K. Vijaya Kumar, How low can you go? Low resolution face recognition study using kernel correlation feature analysis on the FRGCv2 dataset, in: Proceedings of the Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference, 2006, pp. 1–6.
[5] T. Bachmann, Identification of spatially quantised tachistoscopic images of faces: how many pixels does it take to carry identity? Eur. J. Cogn. Psychol. 3 (1) (1991) 87–103.
[6] P. Thevenaz, T. Blu, M. Unser, Interpolation revisited, IEEE Trans. Med. Imaging 19 (7) (2000).
[7] R.G. Keys, Cubic convolution interpolation for digital image processing, IEEE Trans. Acoust. 29 (1981) 1153–1160.
[8] S.W. Lee, J.K. Paik, Image interpolation using adaptive fast B-spline filtering, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 5, 1993, pp. 177–180.
[9] K. Ratakonda, N. Ahuja, POCS based adaptive image magnification, in: Proceedings of the International Conference on Image Processing, 3, 1998, pp. 203–207.
[10] K. Jensen, D. Anastassiou, Subpixel edge localization and the interpolation of still images, IEEE Trans. Image Process. 4 (3) (1995) 285–295.
[11] J.-B. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5197–5206.
[12] L. Wang, Z. Huang, Y. Gong, C. Pan, Ensemble based deep networks for image super-resolution, Pattern Recognit. 68 (2017) 191–198.
[13] J. Zhang, L. Zhang, L. Xiang, Y. Shao, G. Wu, X. Zhou, D. Shen, Q. Wang, Brain atlas fusion from high-thickness diagnostic magnetic resonance images by learning-based super-resolution, Pattern Recognit. 63 (2017) 531–541.
[14] C. Liu, H.-Y. Shum, C.-S. Zhang, A two-step approach to hallucinating faces: global parametric model and local nonparametric model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1, 2001, pp. 192–198.
[15] X. Wang, X. Tang, Hallucinating face by eigentransformation with distortion reduction, in: D. Zhang, A.K. Jain (Eds.), Biometric Authentication, Lecture Notes in Computer Science, 3072, Springer Berlin Heidelberg, 2004, pp. 88–94.
[16] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.
[17] D.D. Lee, H.S. Seung, Learning the parts of objects by nonnegative matrix factorization, Nature 401 (1999) 788–791.
[18] K.I. Kim, Y. Kwon, Single-image super-resolution using sparse regression and natural image prior, IEEE Trans. Pattern Anal. Mach. Intell. 32 (6) (2010) 1127–1133.
[19] X. Ma, J. Zhang, C. Qi, Hallucinating face by position-patch, Pattern Recognit. 43 (6) (2010) 2224–2236.
[20] G. Gao, X.-Y. Jing, P. Huang, Q. Zhou, S. Wu, D. Yue, Locality-constrained double low-rank representation for effective face hallucination, IEEE Access 4 (2016) 8775–8786.
[21] J. Jiang, R. Hu, Z. Wang, Z. Han, Noise robust face hallucination via locality-constrained representation, IEEE Trans. Multimed. 16 (5) (2014) 1268–1281.
[22] J. Jiang, R. Hu, Z. Wang, Z. Han, J. Ma, Facial image hallucination through coupled-layer neighbor embedding, IEEE Trans. Circuits Syst. Video Technol. 26 (9) (2016) 1674–1684.
[23] H. Huang, N. Wu, Fast facial image super-resolution via local linear transformations for resource-limited applications, IEEE Trans. Circuits Syst. Video Technol. 21 (10) (2011) 1363–1377.
[24] X. Zeng, H. Huang, C. Qi, Expanding training data for facial image super-resolution, IEEE Trans. Cybern. 48 (2) (2018) 716–729.
[25] H. Huang, H. He, X. Fan, J. Zhang, Super-resolution of human face image using canonical correlation analysis, Pattern Recognit. 43 (7) (2010) 2532–2543.
[26] L. An, B. Bhanu, Face image super-resolution using 2d cca, Signal Process. 103 (2014) 184–194.
[27] J. Jiang, J. Ma, C. Chen, X. Jiang, Z. Wang, Noise robust face image super-resolution through smooth sparse representation, IEEE Trans. Cybern. 47 (11) (2017) 3991–4002.
[28] A. Akyol, M. GöKmen, Super-resolution reconstruction of faces by enhanced global models of shape and texture, Pattern Recognit. 45 (12) (2012) 4103–4116.
[29] K. Nguyen, C. Fookes, S. Sridharan, M. Tistarelli, M. Nixon, Super-resolution for biometrics: a comprehensive survey, Pattern Recognit. 78 (2018) 23–42.
[30] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 184–199.
[31] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307.
[32] J. Kim, J. Kwon Lee, K. Mu Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1646–1654.
[33] J. Kim, J. Kwon Lee, K. Mu Lee, Deeply-recursive convolutional network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1637–1645.
[34] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: Jul 21 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 105–114.
[35] R. Abiantun, U. Prabhu, M. Savvides, Sparse feature extraction for pose-tolerant face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (10) (2014) 2061–2073.
[36] A.A. Goshtasby, Piecewise linear mapping functions for image registration., Pattern Recognit. 19 (6) (1986) 459–466.
[37] U. Prabhu, J. Heo, M. Savvides, Unconstrained pose-invariant face recognition using 3D generic elastic models, IEEE Trans. Pattern Anal. Mach. Intell. 33 (10) (2011) 1952–1961.
[38] F. Juefei-Xu, K. Luu, M. Savvides, Spartans: single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios, IEEE Trans. Image Process. (TIP) 24 (12) (2015) 4780–4795.
[39] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE, in: Proceedings of the 8th IEEE International Conference on Automatic Face Gesture Recognition, 2008, pp. 1–8.
[40] F. Juefei-Xu, D.K. Pal, K. Singh, M. Savvides, A preliminary investigation on the sensitivity of cots face recognition systems to forensic analyst-style face processing for occlusions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2015, pp. 25–33.
[41] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cogn. Neurosci. 3 (1) (1991) 71–86.
[42] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[43] P.J. Burt, Fast filter transform for image processing, Comput. Graph. Image Process. 16 (1) (1981) 20–51.
[44] S.S. Chen, Basis Pursuit, Stanford University, 1995. Ph.D. thesis.
[45] D.L. Donoho, For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution, Commun. Pure Appl. Math. 59 (2004) 797–829.
[46] E.J. Candes, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Commun. Pure Appl. Math. 59 (8) (2006) 1207–1223.
[47] R.J. Tibshirani, Regression shrinkage and selection via the lasso, J. Royal Stat. Soc. Ser. B 58 (1) (1996) 267–288.
[48] F. Juefei-Xu, M. Savvides, Single face image super-resolution via solo dictionary learning, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 2239–2243.
[49] F. Juefei-Xu, D.K. Pal, M. Savvides, Hallucinating the full face from the periocular region via dimensionally weighted K-SVD, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2014, pp. 1–8.
[50] E.J. Candes, J.K. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inf. Theory 52 (2) (2006) 489–509.

[51] A.Y. Yang, A. Ganesh, Z. Zhou, S. Sastry, Y. Ma, A review of fast ℓ_1-minimization algorithms for robust face recognition, Comput. Res. Repos. abs/1007.3753 (2010).
[52] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 947–954.
[53] P.J. Phillips, S.Z. Der, P.J. Rauss, O.Z. Der, FERET (face recognition technology) recognition algorithm development and test results, Technical Report, ARL-TR-995. Army Research Laboratory, 1996.
[54] F. Juefei-Xu, R. Dey, V.N. Bodetti, M. Savvides, RankGAN: a maximum margin ranking GAN for generating faces, in: Proceedings of the Asian Conference on Computer Vision (ACCV), 2018, pp. 1–14.
[55] F. Juefei-Xu, M. Savvides, Pokerface: partial order keeping and energy repressing method for extreme face illumination normalization, in: Proceedings of the IEEE Seventh International Conference on Biometrics: Theory, Applications, and Systems (BTAS), IEEE, 2015, pp. 1–8.
[56] F. Juefei-Xu, D.K. Pal, M. Savvides, NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2015, pp. 141–150.
[57] F. Juefei-Xu, E. Verma, P. Goel, A. Cherodian, M. Savvides, DeepGender: occlusion and low resolution robust facial gender classification via progressively trained convolutional neural network with attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2016, pp. 68–77.
[58] F. Juefei-Xu, E. Verma, M. Savvides, Deep Learning for Biometrics, Springer, pp. 183–218.
[59] F. Juefei-Xu, M. Savvides, Learning to invert local binary patterns, in: Proceedings of the British Machine Vision Conference (BMVC), BMVA Press, 2016, pp. 1–14.
[60] N. Thomos, N.V. Boulgouris, M.G. Strintzis, Optimized transmission of JPEG2000 streams over wireless channels, IEEE Trans. Image Process. 15 (1) (2006) 54–67.
[61] K. Turkowski, Filters for common resampling tasks, in: Graphics Gems, Academic Press Professional, Inc., 1990, pp. 147–165.
[62] (http://www.ifp.illinois.edu/~jyang29/resources.html).
[63] (http://www.mpi-inf.mpg.de/~kkim/supres/supres.htm).
[64] TensorFlow implementation of SRGAN. https://github.com/tensorlayer/srgan.
[65] PyTorch implementation of SRGAN. https://github.com/leftthomas/SRGAN.
[66] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. (IJCV) 115 (3) (2015) 211–252.
[67] P. Cheeseman, B. Kanefsky, R. Kraft, J. Stutz, R. Hanson, Super-resolved surface reconstruction from multiple images, in: G.R. Heidbreder (Ed.), Maximum Entropy and Bayesian Methods, Fundamental Theories of Physics, 62, Springer Netherlands, 1996, pp. 293–308.
[68] M. Elad, A. Feuer, Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images, IEEE Trans. Image Process. 6 (12) (1997) 1646–1658.
[69] R.C. Hardie, K.J. Barnard, E.E. Armstrong, Joint MAP registration and high-resolution image estimation using a sequence of undersampled images, IEEE Trans. Image Process. 6 (12) (1997) 1621–1633.
[70] R.R. Schultz, R.L. Stevenson, A bayesian approach to image expansion for improved definition, IEEE Trans. Image Process. 3 (3) (1994) 233–242.
[71] F. Juefei-Xu, M. Savvides, An image statistics approach towards efficient and robust refinement for landmarks on facial boundary, in: Proceedings of the IEEE Sixth International Conference on Biometrics: Theory, Applications, and Systems (BTAS), IEEE, 2013, pp. 1–8.
[72] K. Seshadri, F. Juefei-Xu, D.K. Pal, M. Savvides, Driver cell phone usage detection on strategic highway research program (SHRP2) face view videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2015, pp. 35–43.
[73] N. Zehngut, F. Juefei-Xu, R. Bardia, D.K. Pal, C. Bhagavatula, M. Savvides, Investigating the feasibility of image-based nose biometrics, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 522–526.
[74] F. Juefei-Xu, M. Savvides, Can your eyebrows tell me who you are? in: Proceedings of the 5th International Conference on Signal Processing and Communication Systems (ICSPCS), 2011, pp. 1–8.
[75] F. Juefei-Xu, M. Cha, M. Savvides, S. Bedros, J. Trojanova, Robust periocular biometric recognition using multi-level fusion of various local feature extraction techniques, in: Proceedings of the IEEE 17th International Conference on Digital Signal Processing (DSP), IEEE, 2011, pp. 1–7.
[76] F. Juefei-Xu, M. Savvides, Unconstrained periocular biometric acquisition and recognition using cots PTZ camera for uncooperative and non-cooperative subjects, in: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV), IEEE, 2012, pp. 201–208.
[77] S. Venugopalan, F. Juefei-Xu, B. Cowley, M. Savvides, Electromyograph and keystroke dynamics for spoof-resistant biometric authentication, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2015, pp. 109–118.
[78] C. Bhagavatula, C. Zhu, K. Luu, M. Savvides, Faster than real-time facial alignment: a 3D spatial transformer network approach in unconstrained poses, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2, 2017, pp. 3980–3989.
[79] M. O'Brien, PBS NOVA, Manhunt: Boston Bombers, 2013, (http://www.pbs.org/wgbh/nova/tech/manhunt-boston-bombers.html).
[80] M. O'Brien, IMDb for 'PBS NOVA, Manhunt: Boston Bombers', 2013, (http://www.imdb.com/title/tt2945440/).
[81] M. O'Brien, Youtube video for 'PBS NOVA, Manhunt: Boston Bombers', 2013, (https://www.youtube.com/watch?v=uTNDnu1t3LA).
[82] F. Juefei-Xu, C. Bhagavatula, A. Jaech, U. Prasad, M. Savvides, Gait-ID on the move: pace independent human identification using cell phone accelerometer dynamics, in: Proceedings of the IEEE Fifth International Conference on Biometrics: Theory, Applications, and Systems (BTAS), IEEE, 2012, pp. 8–15.
[83] Updates on investigation into multiple explosions in boston, http://www.fbi.gov/news/updates-on-investigation-into-multiple-explosions-in-boston.
[84] Boston marathon bombings - wikipedia, http://en.wikipedia.org/wiki/Boston_Marathon_bombings.
[85] F. Juefei-Xu, M. Savvides, Fastfood dictionary learning for periocular-based full face Hallucination, in: Proceedings of the IEEE Seventh International Conference on Biometrics: Theory, Applications, and Systems (BTAS), IEEE, 2016, pp. 1–8.
[86] M. Savvides, F. Juefei-Xu, U. Prabhu, C. Bhagavatula, Unconstrained biometric identification in real world environments, in: Advances in Human Factors and System Interactions, Springer, Cham. Chicago, 2017, pp. 231–244.
[87] F. Juefei-Xu, M. Savvides, Subspace based discrete transform encoded local binary patterns representations for robust periocular matching on NIST's face recognition grand challenge, IEEE Trans. Image Process. (TIP) 23 (8) (2014) 3490–3505.
[88] P.D. United States Census Bureau, Annual estimates of the resident population for incorporated places over 50,000, ranked by july 1, 2012 population: April 1, 2010 to july 1, 2012, 2012.
[89] C. Xie, M. Savvides, B.V.K. Vijaya Kumar, Redundant class-dependence feature analysis based on correlation filters using Frgc 2.0 data, in: Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), 3, 2005, pp. 153–158.
[90] Face and Iris Evaluation Activities at NIST, http://biometrics.nist.gov/cs_links/face/frgc/FRGC-ICE_Brief_CTST06_post.pdf.
[91] F. Juefei-Xu, V.N. Boddeti, M. Savvides, Local Binary Convolutional Neural Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 19–28.
[92] D.K. Pal, F. Juefei-Xu, M. Savvides, Discriminative invariant kernel features: a bells-and-whistles-free approach to unsupervised face recognition and pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 5590–5599.
[93] F. Juefei-Xu, M. Savvides, Multi-class Fukunaga Koontz discriminant analysis for enhanced face recognition, Pattern Recognit. 52 (2016) 186–205.
[94] F. Juefei-Xu, M. Cha, J.L. Heyman, S. Venugopalan, R. Abiantun, M. Savvides, Robust local binary pattern feature sets for periocular biometric identification, in: Proceedings of the 4th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), IEEE, 2010, pp. 1–8.
[95] F. Juefei-Xu, K. Luu, M. Savvides, T. Bui, C. Suen, Investigating Age invariant face recognition based on periocular biometrics, in: Proceedings of the IEEE/IAPR International Joint Conference on Biometrics (IJCB), IEEE, 2011, pp. 1–7.

Ramzi Abiantun received the Ph.D. degree from Carnegie Mellon University's Electrical and Computer Engineering department in 2013. He was a research assistant at CMU's Cylab Biometric Center from 2006 to 2013, developing computer vision and machine learning techniques for iris and face recognition. His thesis focused on unconstrained face recognition, with an emphasis on pose correction and super-resolution. Prior to enrolling in the Ph.D. program, he completed an Integrated Masters/Bachelor degree also from the Electrical and Computer Engineering department at Carnegie Mellon University, with a concentration in robotics, image processing and pattern recognition.

Felix Juefei-Xu received the Ph.D. degree in Electrical and Computer Engineering (ECE) from Carnegie Mellon University (CMU), Pittsburgh, PA, USA. Prior to that, he received the M.S. degree in ECE and the M.S. degree in Machine Learning from CMU, and the B.S. degree in Electronic Engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China. He is currently with the CMU CyLab Biometrics Center, in a research group specializing in pattern recognition, machine learning, computer vision, and image processing, especially as applied to the field of biometrics. His current research is focused on a fuller understanding of deep learning where he is actively exploring new methods in deep learning that are statistically efficient and adversarially robust. His Ph.D. work is primarily focused on tackling the Pose, Expression, Resolution, Illumination, and Occlusion (Perio) challenges for unconstrained periocular face recognition using shallow and deep discriminative and generative methods, especially under the dome of self-supervised predictive learning. He also has broader interests in pattern recognition, computer vision, machine learning, optimization, statistics, compressive sensing, and image processing. He is the recipient of multiple best/distinguished paper awards, including the Best Poster Paper Award of the IEEE/IAPR International Joint Conference on Biometrics (IJCB) in 2011, the Best Paper Award of the IEEE Seventh International Conference on Biometrics: Theory, Applications and Systems (BTAS) in 2015, the Best Student Paper Award of the IEEE Eighth International Conference on Biometrics: Theory, Applications and Systems (BTAS) in 2016, the ACM SIGSOFT Distinguished Paper Award of the IEEE/ACM International Conference on Automated Software Engineering (ASE) in

2018, and the Best Student Paper Award of the 14th Asian Conference on Computer Vision (ACCV) in 2018.

Utsav Prabhu received the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University (CMU), Pittsburgh, PA, USA in 2015. Prior to that, he received the M.S. degree in Electrical and Computer Engineering and the B.E. degree in Electrical and Telecommunications Engineering from Goa University, India. He worked as a Member of Technical Staff in GS Lab, Pune, India from 2006 to 2008. He is currently a Senior Software Engineer in the Machine Perception team at Google Research. His research interests lie in computer vision and pattern recognition with a focus on understanding humans in video content.

Marios Savvides is the Founder and Director of the Biometrics Center at Carnegie Mellon University and is a Research Professor at the Electrical & Computer Engineering Department and CMU CyLab. He received his B.Eng in Microelectronics Systems Engineering from UMIST, U.K., his Masters of Science in Robotics from the Robotics Institute at Carnegie Mellon University and his PhD from the Electrical and Computer Engineering also at Carnegie Mellon University. He is also one of the tapped researchers to form the Office of the Director of National Intelligence (ODNI) 1st Center of Academic Excellence in Science and Technology (CASIS). His research is mainly focused on developing algorithms for robust face and iris biometrics as well as pattern recognition, machine vision and computer image understanding for enhancing biometric systems performance. He is on the program committee on several Biometric conferences such as IEEE BTAS, ICPR, SPIE Biometric Identification, IEEE AutoID and others as well as organizing and co-chairing Robust Biometrics Understanding the Science & Technology (ROBUST 2008) conference. He was an annual invited speaker at IDGA's main conference on Biometrics for National Security and Defense. He has authored and co-authored over 170 journal and conference publications, including several book chapters in the area of Biometrics and served as the area editor of the Springer's Encyclopedia of Biometrics. He helped co-develop the IEEE Certified Biometrics Professional (CBP) program and was on the main steering committee of the IEEE CBP program. His achievements include leading the R&D in CMU's past participation at NIST's Open Face Recognition Grand Challenge 2005 (CMU ranked #1 in Academia and Industry at hardest experiment #4) and also in NIST's Iris Challenge Evaluation (CMU ranked #1 in Academia and #2 against iris vendors) - his group was the only one to attempt both challenges. Prof. Savvides is listed in Marquis Who's Who in America and in Marquis' Who's Who in Science & Engineering. He has filed over 20 patent applications in area of Biometrics and is the co-recipient of CMU's 2009 Carnegie Institute of Technology (CIT) Outstanding Research Award.