Face Hallucination Using Linear Models of Coupled Sparse Support

Reuben A Farrugia, Christine Guillemot

Face Hallucination using Linear Models of Coupled Sparse Support

Reuben A. Farrugia, *Member, IEEE,* and Christine Guillemot, *Fellow, IEEE*

***Abstract*—Most face super-resolution methods assume that low- and high-resolution manifolds have similar local geometrical structure, hence learn local models on the low-resolution manifold (*e.g*. sparse or locally linear embedding models), which are then applied on the high-resolution manifold. However, the low-resolution manifold is distorted by the one-to-many relationship between low- and high- resolution patches. This paper presents the Linear Model of Coupled Sparse Support (LM-CSS) method which learns linear models based on the local geometrical structure on the high-resolution manifold rather than on the low-resolution manifold. For this, in a first step, the low-resolution patch is used to derive a globally optimal estimate of the high-resolution patch. The approximated solution is shown to be close in Euclidean space to the ground-truth but is generally smooth and lacks the texture details needed by state-of-the-art face recognizers. Unlike existing methods, the sparse support that best estimates the first approximated solution is found on the high-resolution manifold. The derived support is then used to extract the atoms from the coupled low- and high-resolution dictionaries that are most suitable to learn an up-scaling function for every facial region. The proposed solution was also extended to compute face super-resolution of non-frontal images.**

Extensive experimental results conducted on a total of 1830 facial images show that the proposed method outperforms seven face super-resolution and a state-of-the-art cross-resolution face recognition method in terms of both quality and recognition. The best recognition performance was achieved using LM-CSS followed by the Local Binary Pattern (LBP) face recognizer, where it was found to outperform the state-of-the-art Discriminant Face Descriptor (DFD) very-low resolution face recognition system, achieving rank-1 recognition gains between 34% and 60% at very low-resolutions. Moreover, subjective results show that the proposed solution is able to super-resolve more accurate facial images from the challenging IARPA Janus Benchmark A (IJB-A) dataset, which considers a wide range of poses and orientations at magnification factors as high as five.

***Index Terms*—Face hallucination, face recognition, face super-resolution, sparse representation.**

I. Introduction

MOST countries around the world use Closed Circuit Television (CCTV) systems to combat crime in their major cities. These cameras are normally installed to cover a large field of view where the query face image may not be sampled densely enough by the camera sensors [1]. The low-resolution and quality of face images captured on camera reduces the effectiveness of CCTV in identifying perpetrators and potential eyewitnesses [2], [3].

Manuscript received November 12, 2015;

R.A.Farrigoa is with the Department of Communications and Computer Engineering, University of Malta, Malta e-mail: (reuben.farrugia@um.edu.mt).

C. Guillemot is with the Institut National de Recherche en Informatique et en Automatique, Rennes 35042, France (christine.guillemot@inria.fr).

Super-resolution techniques can be used to enhance the quality of low-resolution facial images to improve the recognition performance of existing face recognition software and the identification of individuals from CCTV images. In a recent survey Wang *et al*. distinguish between two main categories of super-resolution methods: reconstruction based and learning based approaches [4]. Reconstruction based methods register a sequence of low-resolution images onto a high-resolution grid and fuse them to suppress the aliasing caused by under-sampling [5], [6]. On the other hand, learning based methods use coupled dictionaries to learn the mapping relations between low- and high- resolution image pairs to synthesize high-resolution images from low-resolution ones [7], [8]. The research community has lately focused on the latter category of super-resolution methods, since they can provide higher quality images and larger magnification factors. Generic super-resolution techniques [7], [8] can be used to super-resolve facial images. Recent advances in this area have proposed to model the up-scaling function of generic images using machine learning [9], [10].

In their seminal work, Baker and Kanade [11] exploited the fact that human face images are a relatively small subset of natural scenes and introduced the concept of face super-resolution (also known as face hallucination) where only facial images are used to construct the dictionaries. The high-resolution face image is then hallucinated using Bayesian inference with gradient priors. The authors in [12] assume that two similar face images share similar local pixel structures so that each pixel could be generated by a linear combination of spatially neighbouring pixels. This method was later extended in [13] where they use sparse local pixel structure. Although these methods were found to perform well at moderately low-resolutions, they fail when considering very low-resolution face images where the local pixel structure is severely distorted. Classical face representation models were used to model a novel low-resolution face image using a linear combination of prototype low-resolution face images present in a dictionary [14]–[21]. The combination weights are then used to combine the corresponding high-resolution prototype face images to hallucinate the high-resolution face image. Nevertheless, global methods do not manage to recover the local texture details which are essential for face recognition.

Data representation methods have also been used to hallucinate high-resolution overlapping patches which are then stitched together to reconstruct the high-resolution face image [22]–[39]. Post processing techniques were used in [16]–[18], [20]–[26] to recover more texture detail. Nevertheless, both global and local methods [14]–[25], [27]–[39] assume

that low- and high-resolution manifolds have similar local geometrical structures. However, the authors in [28], [40]–[42] have shown that this assumption does not hold well because the one-to-many mappings between low- and high-resolution patches distort the structure of the low-resolution manifold. Therefore, the reconstruction weights estimated on the low-resolution manifold do not correlate well with the actual weights needed to reconstruct the unknown high-resolution patch on the high-resolution manifold.

Motivated by this observation, the authors in [41]–[44] derive a pair of projection matrices that can be used to project both low- and high-resolution patches on a common coherent subspace. However, the dimension of the coherent sub-spaces is equal to the lowest rank of the low- and high-resolution dictionary matrices. Therefore, the projection from the coherent sub-space to the high-resolution manifold is ill-conditioned. On the other hand, the Locality-constrained Iterative Neighbour Embedding (LINE) method presented in [45], [46] reduces the dependence from the low-resolution manifold by iteratively updating the neighbours on the high-resolution manifold. This was later extended by the same authors in [47] where an iterative dictionary learning scheme was integrated to bridge the low- and high-resolution manifolds. Although this method yields state-of-the-art performance, it cannot guarantee to converge to an optimal solution. A recent method based on Transport Analysis was proposed in [48] where the high-resolution face image is reconstructed by morphing high resolution training images which best fit the given low-resolution face image. However, this method heavily relies on the assumption that the degradation function is known, which is generally not possible in typical CCTV scenarios. An ensemble of feature-based regression functions and classifiers was used in [49], while deep learning was used in [50] for face de-blurring applications. While the latter approach claims to perform hallucination in the wild, the tests were conducted on faces which are almost frontal.

Different automated cross-resolution face recognition methods have been proposed to cater for the resolution discrepancy between the gallery and probe images[1]. These methods either try to include the resolution discrepancy within the classifier's optimization function [1], [51]–[53] or else by projecting both probe and gallery images on a coherent subspace and compute the classification there [54]–[58]. However, although these methods are reported to provide good results, they suffer from the following shortcomings i) most of the methods ([52], [54]–[58]) do not synthesize a high resolution face image (unlike face hallucination methods) and ii) they generally assume that several images of the same subject are available in the gallery, which are often scarce in practice.

This work presents a two layer approach named Linear Models of Coupled Sparse Support (LM-CSS), which employs a coupled dictionary containing low- and high-resolution training patches to learn the optimal up-scaling function for each patch corresponding to a specific facial region. The main contributions of this work are summarized below:

1) We show experimentally (see section III) that learning mapping relations that maximize the Peak Signal-to-Noise Ratio (PSNR) quality metric (as done by existing face hallucination methods) generally provides smooth facial images that lack texture details essential for face recognition and person identification. We also demonstrate that more texture consistent solutions can be derived when using a sparse subset of atoms from a over-complete coupled dictionaries.
2) The LM-CSS method described in section IV is a novel two-stage face hallucination method that differs from existing methods:
    a) All existing face hallucination methods discussed above [14]–[25], [27]–[39], [41]–[48] synthesize the high resolution test image using a weighted combination of high-resolution face images. The resulting synthesized face is not accepted as criminal evidence in court since it morphs a number of faces. On the other hand, both stages of LM-CSS use the facial images contained within coupled dictionaries to learn the optimal up-scaling function for each patch. The design philosophy of LM-CSS is close to interpolation schemes (modelling an up-scaling function) which are well accepted to restore forensic evidence.
    b) Opposed to existing methods [14]–[25], [27]–[39], [41]–[44], LM-CSS exploits the geometrical structure of the high-resolution manifold, which is not distorted, to select the atoms that are most suitable to learn the the up-scaling function for each patch. We show experimentally in Section III that the first stage of LM-CSS derives a first approximation which better preserves the local neighbourhood. The second stage finds the sparse coupled support to be used to estimate the final up-scaling function which is able to reconstruct facial images which are more coherent to the ground-truth.
    c) Unlike the methods in [45]–[47] the proposed method is non-iterative and is guaranteed to converge to an optimal solution.
3) Existing face hallucination methods are designed and tested on frontal face images. This limits their use in practice since facial images captured by CCTV are often far from frontal pose. This work proposes a method that exploits facial landmark points to align the coupled dictionaries with the input low-resolution face. This method can be applied to all face hallucination schemes, including LM-CSS (see Section VI-D). While being simple, this can be considered as the first attempt to extend existing face hallucination methods to perform on facial images whose orientation vary significantly from the frontal pose.

Apart from this, we demonstrate that while existing face hallucination schemes try to derive a solution that minimizes the mean square error (MSE) and assume that it will improve

[1]Gallery images are high-quality frontal facial images stored in a database which are usually taken in a controlled environment (e.g. ID and passport photos). Probe images are query face images which are compared to each and every face image included in the gallery. Probe images are usually taken in a non-controlled environment and can have different resolution.

recognition, we show here that this is in fact not true. The MSE metric is biased toward blurred images [66] that lack texture detail essential for person identification and recognition. In fact, the analysis in Section III and results in Section VI show that reconstructing faces whose texture is more coherent to the ground truth aids the recognition performance more than solutions that minimize the MSE. These findings are in line with recent results obtained in the related field of face sketch recognition [59].

The proposed approach has been extensively evaluated against seven face hallucination and one cross-resolution face recognition methods using 930 probe images from the FRGC dataset against a gallery of 889 individuals and 900 probe images from the CAS-PEAL dataset against a gallery of 1040 individuals using a closed set identification scenario with one face image per subject in the gallery[2]. The quality analysis was conducted using all 1830 probe images from both datasets. The best rank-1 recognition performance was attained using the proposed LM-CSS face super-resolution and using the LBP face recognizer which achieved rank-1 recognition gains between 34% and 60% over the Discriminant Face Descriptor (DFD) [58] cross-resolution face recognizer at very low resolutions, around 1% gain over the most competitive method LINE+LBP and between 2% and 8% over Eigen-Patches+LBP. The quality analysis further shows that the proposed method is competitive, and most of the time superior to Eigen-Patches while it outperforms LINE by around 1dB in Peak Signal-to-Noise Ratio (PSNR). Subjective results further demonstrate that the proposed LM-CSS can be used to super-resolve facial images from the challenging IARPA Janus Benchmark A (IJB-A) dataset and generates facial images with more texture detail and lower visual distortions.

The rest of the paper is organized as follows. After introducing the notations in Section II, we analyse the Neighbour Embedding scheme of [61] which is the basis of the most successful schemes in face hallucination from which we derive the observations on which our method will be based on. The proposed LM-CSS method is described in Section IV while the proposed extension to perform face super-resolution in the wild are provided in Section V. The testing methodology and results are provided in Section VI while the final concluded remarks are provided in Section VII.

II. Problem Formulation

We consider a low-resolution face image $\mathbf{X}$ where the distance between the eye centres (inter-eye distance) is defined as d_x. The goal of face super-resolution is to up-scale $\mathbf{X}$ by a scale factor $\alpha = \frac{d_y}{d_x}$, where d_y represents the inter-eye distance of the desired super-resolved face image. The image $\mathbf{X}$ is divided into a set of overlapping patches of size $\sqrt{n} \times \sqrt{n}$ with an overlap of γ_x, and the resulting patches are reshaped to column-vectors in lexicological order and stored as vectors $\mathbf{x}_i$, where $i \in [1, p]$ represents the patch index.

[2]Collection of gallery images is laborious and expensive in practice. This limits the number of gallery images that can be used in practice for recognition, where frequently only one image per subject is available in the gallery. This problem is referred to as the one sample per person in face recognition literature [60].

In order to learn the up-scaling function between low- and high-resolution patches, we have m high resolution face images which are registered based on eye and mouth center coordinates, where the inter-eye distance is set to d_y. These images are divided into overlapping patches of size $[\alpha\sqrt{n} \times \alpha\sqrt{n}]$ with an overlap of $\gamma_y = [\alpha\gamma_x]$, where $[*]$ stands for the rounding operator. The i-th patch of every high-resolution image is reshaped to column-vectors in lexicological order and placed within the high-resolution dictionary $\mathbf{H}_i$. The low-resolution dictionary of the i-th patch $\mathbf{L}_i$ is constructed using the same images present in the high-resolution dictionary, which are down-scaled by a factor $\frac{1}{\alpha}$ and divided into overlapping patches of size $\sqrt{n} \times \sqrt{n}$ with an overlap of γ_x. This formulation is in line with the position-patch method published in [33] where only collocated patches with index i are used to super-resolve the low resolution patch $\mathbf{x}_i$.

Without loss of generalization we will assume that the column vectors of both dictionaries are standardized to have zero mean and unit variance to compensate for illumination and contrast variations. The standardized low-resolution patch is denoted by $\mathbf{x}_i^s$ and the aim of this work is to find an up-scaling projection matrix that minimizes the following objective function

$$\mathbf{\Phi}_i = \arg\min_{\mathbf{\Phi}_i} ||\mathbf{H}_i - \mathbf{\Phi}_i \mathbf{L}_i||_2^2 \tag{1}$$

where $\mathbf{\Phi}_i$ is the up-scaling projection matrix of dimensions $[\alpha]^2 n \times n$. The standardized i-th high-resolution patch is then hallucinated using

$$\widetilde{\mathbf{y}}_i^s = \mathbf{\Phi}_i \mathbf{x}_i^s \tag{2}$$

where $\mathbf{x}_i^s$ is the standardized i-th low-resolution patch. In the sequel, the upper-script s indicates that the vector is standardized to have zero mean and unit variance. The pixel intensities of the patch are then recovered using

$$\widetilde{\mathbf{y}}_i = [\sigma_i \widetilde{\mathbf{y}}_i^s + \mu_i] \tag{3}$$

where μ_i and σ_i represent the mean and standard deviation of the low-resolution patch $\mathbf{x}_i$. The resulting hallucinated patches are then stitched together by averaging overlapping pixels to form the hallucinated high-resolution face image $\widetilde{\mathbf{Y}}$.

This formulation is considerably different from the one commonly used in [14]–[25], [27]–[39], [41]–[48] where they try to find the optimal reconstruction weights $\boldsymbol{w}_i$ that minimize the following optimization function

$$\boldsymbol{w}_i = \arg\min_{\boldsymbol{w}_i} ||\mathbf{x}_i^s - \mathbf{L}_i \boldsymbol{w}_i||_2^2 \tag{4}$$

and applying some additional regularization terms to improve the approximation. In essence, these methods exploit the structure of the low-resolution manifold to find the optimal combination of atoms from the low-resolution dictionary that best reconstructs the low-resolution test patch $\mathbf{x}_i^s$. The same combination weights are then used to synthesize the high-resolution patch using

$$\widetilde{\mathbf{y}}_i^s = \mathbf{H}_i \boldsymbol{w}_i \tag{5}$$

which is a weighted combination of the atoms contained within the high-resolution dictionary.

III. Quality and Texture Analysis

Motivated by the success of Neighbour Embedding (NE) [61], which forms the basis of existing state-of-the-art face hallucination schemes, we investigate here the effect that the number of atoms (or neighbours) has on its performance. We emphasize here that the results presented in this section are computed using the NE algorithm, and not the LM-CSS proposed in this paper which will be described in Section IV. In this experiment we consider NE and we vary the number of neighbours k used to compute the weighted combination. Figure 1 depicts the quality and texture analysis as a function of k using different magnification factors. These results were computed using all the 886 images from the AR face dataset [62] while the coupled dictionary is constructed using one-image per subject from the Color Feret [63] and Multi-Pie [64] datasets as described in section VI-A. The quality was measured using PSNR[3] while the Texture Consistency (TC) was measured by comparing the LBP features of the reference and hallucinated images. The LBP features were extracted using the method in [65] where the similarity was measured using histogram intersection. In this experiment $n = 25$, $\gamma_x = 2$ and $m = 1203$.

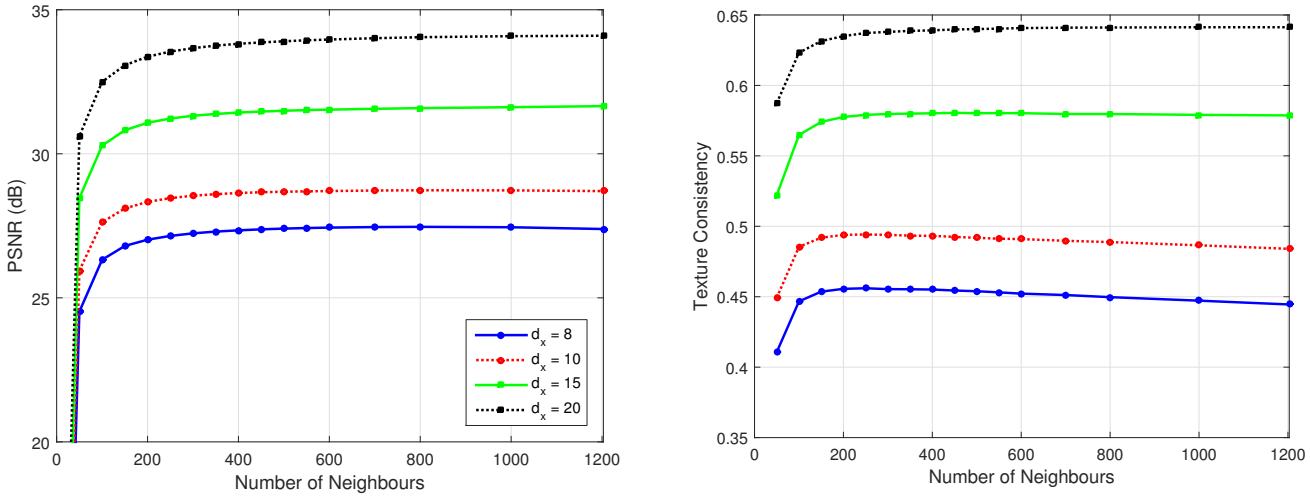
The results in Figure 1a demonstrate that the PSNR increases rapidly until $k = 200$, and keeps on improving slowly for larger values of k. The highest PSNR value was obtained when $k = m$ *i.e.* all column-vectors are used to approximate the optimal combination weights. However, the results in Figure 1b show that the texture consistency increases steadily up till $k = 200$ and starts degrading (or remains steady) as k increases. This indicates that the texture between the reference and hallucinated image is more consistent when using a small number of atoms (*i.e.* $k = 200$) while a larger neighbourhood size will provide blurred images which lack important texture details. The subjective results in Figure 2 support this observation where it can be seen that the images derived using $k = 200$ generally contain more texture details while the images for $k = m = 1203$, which attain larger PSNR values, are more blurred. We also present the spectrum of the LBP texture descriptor to see why the texture consistency decreases when $k > 200$. It can be seen that the LBP spectrum using NE with $k = 200$ is closer to the spectrum of the high-resolution facial images, which is confirmed by the higher TC metric. Moreover, one can observe that the LBP spectrum of NE using $k = 1203$ contains more noisy spikes and is more sparse *i.e.* less non-zero coefficients. These spikes can be explained by the fact that since the face images restored using $k = 1203$ are blurred, they contain more repetitive texture, and therefore the energy of the spectrum is contained within a smaller number of coefficients.

All the face hallucination methods found in literature [14]–[25], [27]–[39], [41]–[48] follow the same philosophy of generic super-resolution and are designed to maximize an objective measure such as PSNR. In spite of that, the PSNR quality metric depends on the squared difference between the distorted and original image in a holistic manner and is biased to favour blurred facial images which is inconsistent with the human vision system (HVS) [66]. All these methods assume that increasing the PSNR will inherently improve the face recognition performance. The above results and observations reveal that improving the PSNR does not correspond to improving the texture detail of the hallucinated face image. Moreover, state-of-the-art face recognition methods [65], [67]–[69] exploit the texture similarity between probe and gallery images to perform automated facial recognition. This indicates that optimizing the face hallucination to minimize the mean square error leads to sub-optimal solutions, at least in terms of recognition. Therefore, comparing face hallucination methods using solely the PSNR quality metric (as done by all papers on face hallucination) provide misleading conclusions on which method performs best, since it ignores the texture consistency between the reference The results in Fig. 2 and remarks in the related field of face scatch recognition [59] shows that it is more important to recover texture detail coherent with the reference face image than reducing the mean square error, since recognition and identification exploits facial texture to discriminate between different individuals. The results in Fig. 1b further show that there is a relation between texture similarity and sparsity, i.e. facial images hallucinated using the k-nearest atoms, where $(k \ll m)$, are more consistent in terms of texture. This observation is used in the design of the proposed LM-CSS which will be described in Section IV.

[3]Similar results were obtained using other full-reference quality metrics such as Structural Similarity (SSIM) and Feature Similarity (FSIM) metrics.

IV. Linear Models of Coupled Sparse Support

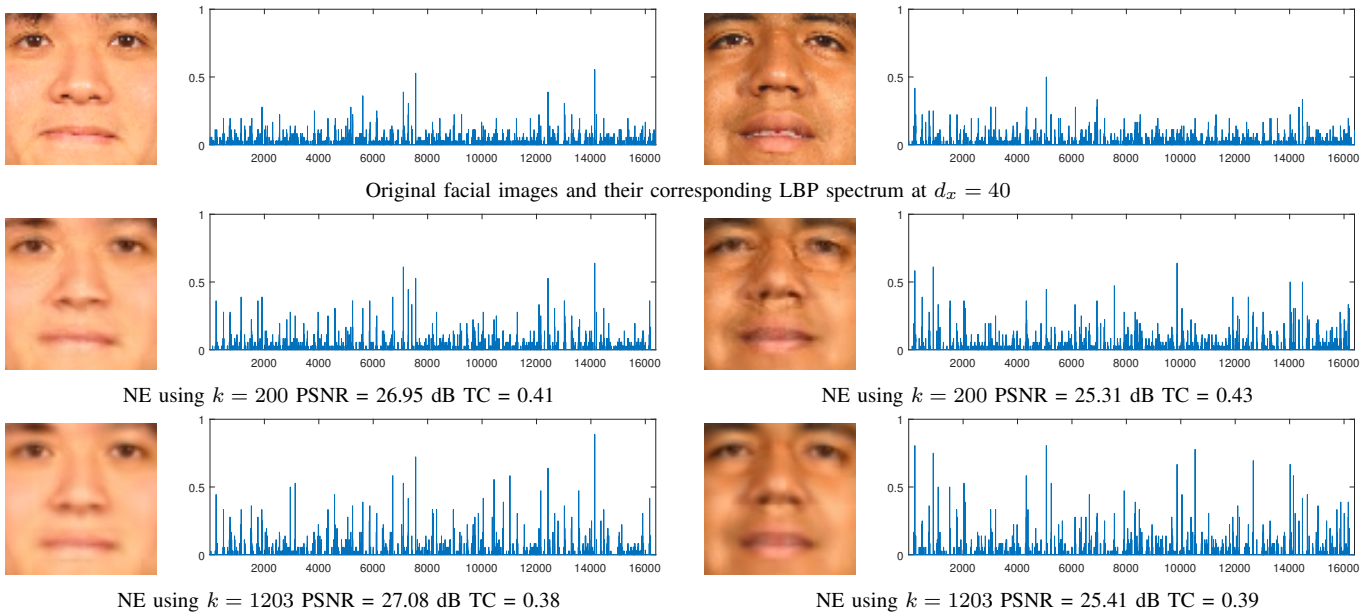
The proposed method builds on the observations drawn in the previous section where the main objective is to find the atoms that are able to better preserve the similarity between the hallucinated and ground-truth images in terms of both texture and quality. A schematic diagram of the proposed method is shown in Fig. 3. The aim of this approach is to learn an up-scaling function for each patch by exploiting the local geometrical structure of the high-resolution manifold. Fig. 3 shows the block-diagram of the proposed LM-CSS method, where in this example the first patch ($i = 1$) covering the right eye is being processed. The low-resolution patch $\mathbf{x}_i$ is first standardized (zero mean and unit variance) and then passed to the first layer (*Layer 0*) which derives the first approximation $\widetilde{\mathbf{y}}_i^{s\{0\}}$ by modelling an up-scaling function $\mathbf{\Phi}_i^{\{0\}}$ for every i-th patch using all the elements (or atoms) within the coupled patch dictionaries $\mathbf{L}_i$ and $\mathbf{H}_i$. This solution is seen here as a point on the high-resolution manifold which is closest to the ground-truth in Euclidean space. While this solution is optimal in terms of Euclidean distance, it lacks texture detail which is essential for face identification and recognition. Given that the first approximated solution $\widetilde{\mathbf{y}}_i^{s\{0\}}$ is sufficiently close to the ground-truth (which is unknown), $\widetilde{\mathbf{y}}_i^{s\{0\}}$ will share similar local structure on the high-resolution manifold, which was proved to be valid in other research domains [70], [71]. The purpose of the second layer (*Layer 1*) is then to exploit this similarity between the local structure of $\widetilde{\mathbf{y}}_i^{s\{0\}}$ and



(a) Quality Analysis

(b) Texture Analysis

Fig. 1: Performance analysis using different neighbourhood size k.



Original facial images and their corresponding LBP spectrum at $d_x = 40$

NE using $k = 200$ PSNR = 26.95 dB TC = 0.41

NE using $k = 200$ PSNR = 25.31 dB TC = 0.43

NE using $k = 1203$ PSNR = 27.08 dB TC = 0.38

NE using $k = 1203$ PSNR = 25.41 dB TC = 0.39

Fig. 2: The upper row shows the original face image and corresponding LBP spectrums, and the second and third rows show the super-resolved faces and corresponding LBP spectrum using NE with $k = 200$ and $k = 1203$ respectively

the ground truth to derive the atoms which are optimal to reconstruct $\widetilde{\mathbf{y}}_i^{s\{0\}}$, and we define the sparse weights with the largest k coefficients as the coupled sparse support. However, instead of using the sparse weight coefficients directly, which will approximate the blurred first solution $\widetilde{\mathbf{y}}_i^{s\{0\}}$, we use the coupled sparse support to select the corresponding atoms from the dictionaries $\mathbf{L}_i$ and $\mathbf{H}_i$ to derive a refined up-scaling function $\mathbf{\Phi}_i^{\{1\}}$ which is able to preserve better the texture of the reconstructed i-th patch. In the next subsections we detail the contribution of both layers.

A. *Layer 0: First Approximated Solution*

Driven by the observations illustrated in Section III, the aim of *Layer 0* is to derive an up-scaling function which solves the following L_2-regularized least squares problem

$$\mathbf{\Phi}_i^{\{0\}} = \arg\min_{\mathbf{\Phi}_i^{\{0\}}} ||\mathbf{H}_i - \mathbf{\Phi}_i^{\{0\}}\mathbf{L}_i||_2^2 \text{ subject to } ||\mathbf{\Phi}_i^{\{0\}}||_2^2 \leq \delta^{\{0\}} \tag{6}$$

where all the atoms in dictionaries $\mathbf{L}_i$ and $\mathbf{H}_i$ are used to model the up-scaling function for the i-th patch. This has a closed form solution given by

$$\mathbf{\Phi}_i^{\{0\}} = \mathbf{H}_i\mathbf{L}_i^T\left(\mathbf{L}_i\mathbf{L}_i^T + \lambda^{\{0\}}\mathbf{I}\right)^{-1} \tag{7}$$

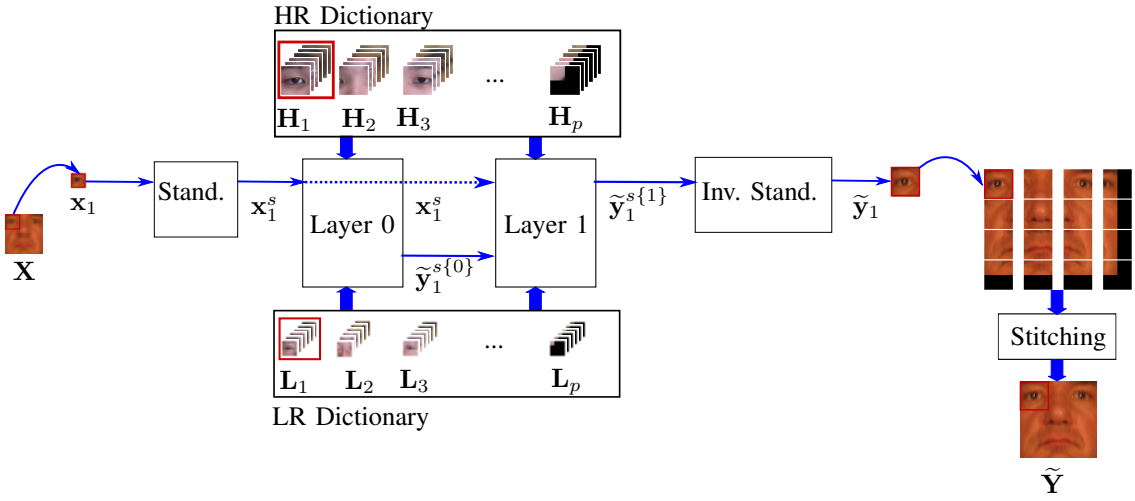


Fig. 3: Block Diagram of the proposed method.

where $\lambda^{\{0\}}$ is a regularization parameter which ensures that the covariance matrix $\mathbf{L}_i\mathbf{L}_i^T$ is invertible and $\mathbf{I}$ is the identity matrix. This solution ignores the local geometric structure of the low-resolution manifold which is known to be distorted, and approximates the upscaling function using the global structure of the low-resolution examples included in $\mathbf{L}_i$. The first approximation $\widetilde{\mathbf{y}}_i^{s\{0\}}$ is then computed using

$$\widetilde{\mathbf{y}}_i^{s\{0\}} = \mathbf{\Phi}_i^{\{0\}}\mathbf{x}_i^s \tag{8}$$

This provides a unique and global solution for the approximation of the ground-truth. Backed up by the results in Fig. 1a, this solution, which employs all elements within the coupled dictionaries, provides the largest PSNR and is thus close to the ground-truth in Euclidean space.

In order to characterize the locality of the proposed first approximation $\widetilde{\mathbf{y}}_i^{s\{0\}}$ with respect to the ground-truth, we use the neighbourhood preservation metric which is defined by

$$np_{L,i,k} = \frac{1}{k}\left(NN(\mathbf{y}_i^s, \mathbf{H}_i, k) \cap NN(\mathbf{x}_i^s, \mathbf{L}_i, k)\right) \tag{9}$$

$$np_{H,i,k} = \frac{1}{k}\left(NN(\mathbf{y}_i^s, \mathbf{H}_i, k) \cap NN(\widetilde{\mathbf{y}}_i^{s\{0\}}, \mathbf{H}_i, k)\right) \tag{10}$$

where $np_{L,i,k}$ and $np_{H,i,k}$ corresponding to the neighbour preservation when searching for neighbours on the low- and high-resolution dictionaries respectively. Here, the function $NN(\mathbf{x}, \mathbf{D}, k)$ derives the k-nearest neighbours of vector $\mathbf{x}$ which minimize the MSE from the dictionary $\mathbf{D}$. The results in Fig. 4 clearly demonstrates that searching on the high-resolution manifold using the up-scaling function $\mathbf{\Phi}_i^{\{0\}}$ computed using *Layer 0* is more beneficial than searching on the low-resolution manifold, since one can find neighbourhoods more coherent with the ground truth at different neighbourhood sizes. We emphasize here that all face hallucination techniques search for the neighbours on the low-resolution manifold, except for the LINE method [47], where they start by searching on the low-resolution manifold and then try to refine the neighbourhood search iteratively on the high resolution manifold. Nevertheless, the LINE method is not guaranteed to converge while our first approximation is computed using a closed form solution.

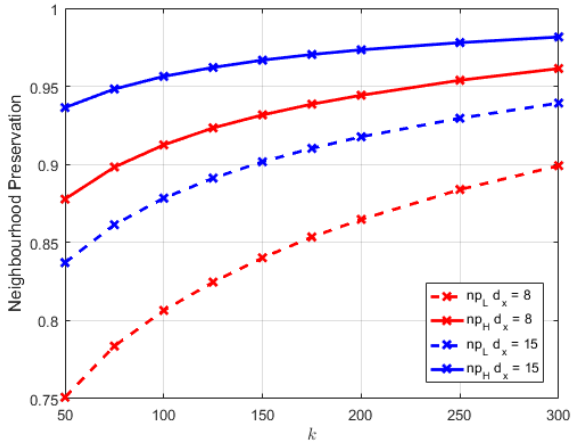


Fig. 4: Neighbour preservation computed using all images from the AR dataset averaged over all patches and images. The solid lines depict the performance when searching in the high-resolution dictionary $\mathbf{H}_i$ using the first approximation $\widetilde{\mathbf{y}}_i^{s\{0\}}$ while the dotted lines indicate nearest neighbourhood searches on the low-resolution dictionary $\mathbf{L}_i$.

Searching for neighbours on the high-resolution manifold results in improving the quality of the reconstructed patches. Fig. 5 compares the classical Neighbour Embedding which derives the k neighbours on the low resolution manifold (NE-LRM) with an extended version of Neighbour Embedding where the k-nearest neighbours of $\widetilde{\mathbf{y}}_i^{s\{0\}}$ on the high-resolution manifold are used to derive the optimal reconstruction weights. These results clearly show that searching for neighbours on the high-resolution manifold improves the neighbourhood preservation and results in improving the quality of the reconstructed patches by achieving lower root MSE (RMSE). Nevertheless, the up-scaling function $\mathbf{\Phi}_i^{\{0\}}$ which employs all the elements within the coupled dictionary, generates facial images which are blurred and lack important texture details which reduces

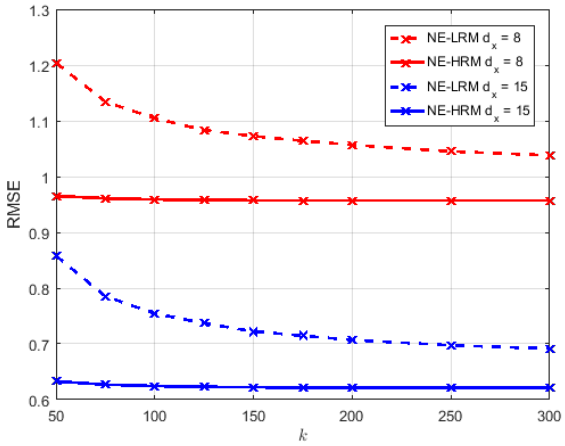


Fig. 5: RMSE analysis comparing NE-LRM and NE-HRM at different resolutions.

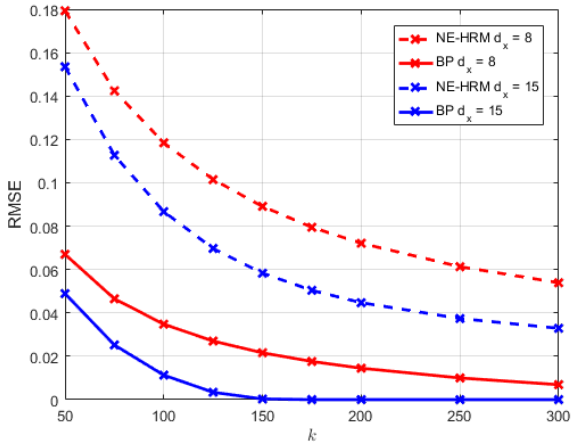


Fig. 6: RMSE showing the error between the $\widetilde{\mathbf{y}}_i^{s\{0\}}$ and those modelled using NE-HRM and BP. Both methods seek for the support $\mathbf{s}_i$ on the high-resolution manifold by exploiting the locality of $\widetilde{\mathbf{y}}_i^{s\{0\}}$. The solid lines depict the performance using BP while the dotted lines show the RMSE obtained nusing NE-HRM at different resolutions.

the discrimination between different faces (see results in Section III).

B. Layer 1: Refined up-scaing function

The aim of *Layer 1* is to find the corresponding coupled column-vectors within the coupled dictionaries $\mathbf{L}_i$ and $\mathbf{H}_i$, which will be referred to as coupled sparse support $\mathbf{s}_i$ for the i-th patch that best reconstruct the first approximated solution $\widetilde{\mathbf{y}}_i^{s\{0\}}$. One simple solution could be to find the k-nearest neighbours of $\widetilde{\mathbf{y}}_i^{s\{0\}}$ and then use Locally Linear Embedding (LLE) [72] as done in [61] to reconstruct it. While this approach exploits the locality of the first approximation to find the support, it does not ensure that the optimal atoms are selected. In this work we formulate this problem to minimize the following objective function

$$\begin{aligned} \boldsymbol{\eta}_i &= \arg\min_{\boldsymbol{\eta}_i} ||\boldsymbol{\eta}_i||_0 = k \\ \text{subject to } &||\widetilde{\mathbf{y}}_i^{s\{0\}} - \mathbf{H}_i\boldsymbol{\eta}_i||_2^2 \leq \delta^{\{1\}} \end{aligned} \tag{11}$$

where $\boldsymbol{\eta}_i$ is the sparse vector, $||\boldsymbol{\eta}_i||_0$ represents the number of non-zero entries in $\boldsymbol{\eta}_i$ which is constrained to be equal to k and $\delta^{\{1\}}$ is the noise parameter. This optimization seeks for the k atoms in $\mathbf{H}_i$ that are most suitable to reconstruct $\widetilde{\mathbf{y}}_i^{s\{0\}}$. The authors in [73], [74] have shown that (11) can be relaxed and solved using Basis Pursuit Denoising (BPDN) which is formulated by

$$\boldsymbol{\eta}_i = \arg\min_{\boldsymbol{\eta}_i} ||\widetilde{\mathbf{y}}_i^{s\{0\}} - \mathbf{H}_i\boldsymbol{\eta}_i||_2^2 + \lambda_s||\boldsymbol{\eta}_i||_1 \tag{12}$$

where λ_s is a regularization parameter. This optimization can be solved in polynomial time using linear programming. In this work we use the solver provided by SparseLab[4] to solve the above BPDN problem. The support $\mathbf{s}_i$ is then set as the k indices of $\boldsymbol{\eta}_i$ with the largest magnitude. The results in Fig. 6 show that modelling $\widetilde{\mathbf{y}}_i^{s\{0\}}$ using BP provides significantly better performance in terms of RMSE when compared to the NE-HRM method described in the previous subsection. This shows that BPDN can find the optimal support $\mathbf{s}_i$ and can represent $\widetilde{\mathbf{y}}_i^{s\{0\}}$ with only k atoms.

Fig. 7 depicts the geometrical representation of the the proposed method. The first approximated solution $\widetilde{\mathbf{y}}_i^{s\{0\}}$ is sufficiently close to the ground-truth, which is generally not known. Nevertheless, *Layer 1* exploits the local-structure on the high-resolution manifold to find the k-column vectors whose support corresponds to the indices of the k non-zero coefficients of $\boldsymbol{\eta}_i$ with the largest magnitude. Given that $\widetilde{\mathbf{y}}_i^{s\{0\}}$ is sufficiently close to the ground truth, we assume that the optimal support $\mathbf{s}_i$ suitable to reconstruct $\widetilde{\mathbf{y}}_i^{s\{0\}}$ is a good approximation of the actual support of the ground-truth. This assumption is supported by empirical results in [71] where they demonstrated that the sparse vector $\boldsymbol{\eta}_i$ tends to be local, *i.e.* the support of two vectors that are sufficiently close are relatively correlated.

Nevertheless, it is important to emphasize here that reconstructing the i-th patch using the weighted combination $\mathbf{H}_i\boldsymbol{\eta}_i$ will provide a solution very close to the approximated solution $\widetilde{\mathbf{y}}_i^{s\{0\}}$ and will therefore still lack texture details. Instead, we define two coupled sub-dictionaries $\mathbf{L}_i(\mathbf{s}_i)$ and $\mathbf{H}_i(\mathbf{s}_i)$, which correspond to the atoms marked in orange in Fig. 7, and we use them to derive a projection matrix $\boldsymbol{\Phi}_i^{\{1\}}$ which minimizes the following objective function

$$\begin{aligned} \boldsymbol{\Phi}_i^{\{1\}} &= \arg\min_{\boldsymbol{\Phi}_i^{\{1\}}} ||\mathbf{H}_i(\mathbf{s}_i) - \boldsymbol{\Phi}_i^{\{1\}}\mathbf{L}_i(\mathbf{s}_i)||_2^2 \\ &\text{subject to } ||\boldsymbol{\Phi}_i^{\{1\}}||_2^2 \leq \delta^{\{1\}} \end{aligned} \tag{13}$$

which has a closed form solution given by

[4]The code can be found at https://sparselab.stanford.edu/

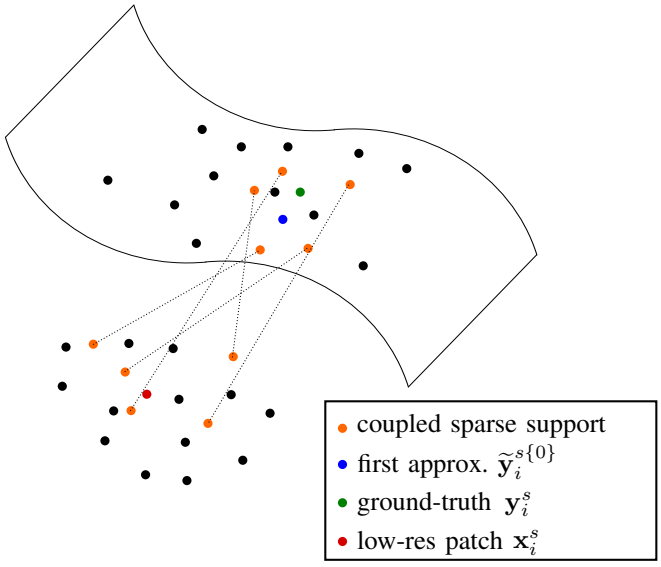


Fig. 7: Geometrical representation of the proposed LM-CSS method.

$$\mathbf{\Phi}_i^{\{1\}} = \mathbf{H}_i(\mathbf{s}_i)\mathbf{L}_i(\mathbf{s}_i)^T \left(\mathbf{L}_i(\mathbf{s}_i)\mathbf{L}_i(\mathbf{s}_i)^T + \lambda^{\{1\}}\mathbf{I}\right)^{-1} \quad (14)$$

where $\lambda^{\{1\}}$ is a regularization parameter and $\mathbf{I}$ is the identity matrix. The refined projection matrix employs only k atoms and is therefore expected to provide a solution which is less smooth and thus preserve the texture details important for recognition. On the other hand, given that the support $\mathbf{s}_i$ is optimal to reconstruct the initial solution $\widetilde{\mathbf{y}}_i^{s\{0\}}$, we expect that improving the texture consistency will not significantly reduce the PSNR quality metric. The resulting super-resolved standardized patch is then approximated using

$$\widetilde{\mathbf{y}}_i^{s\{1\}} = \mathbf{\Phi}_i^{\{1\}}\mathbf{x}_i^s \quad (15)$$

The resulting hallucinated patch $\widetilde{\mathbf{y}}_i^{s\{1\}}$ is then inverse standardized using (3) to recover the actual pixel intensities. The last step involves stitching the overlapping patches together, which is computed by averaging overlapping pixels.

The complexity of the first layer which employs all m atoms in the dictionary is of order $O(n^3)$. The second layer first computes BPDN followed by Multivariate Ridge Regression on the selected k support points. In this work we use the SparseLab solver for BPDN which employs Primal-Dual Interior-Point Algorithm whose complexity is of order $O(m^3)$. The complexity of Multivariate Ridge Regression using k support vectors is of the order $O(n^3)$. This analysis reveals that the complexity of the proposed method is mostly dependent on the sparse solver used, where existing state-of-the-art solvers can reduce it by orders of magnitude than $O(m^3)$ [76].

V. Face Hallucination in the Wild

Existing face hallucination methods are only suitable to super-resolve frontal facial images. The main problem is that the coupled dictionaries contain frontal face images and are therefore not suitable to compute face hallucination in the wild. This section presents a simple, yet effective method that registers the coupled dictionaries to the orientation of the face image being processed. Fig. 8 shows a schematic diagram of the proposed method. The landmark points are manually marked on the input low-resolution test image and the coordinates of each landmark point are stored in $\mathbf{z}_x$. The high-resolution landmark points are then approximated using $\mathbf{z}_y = \alpha\mathbf{z}_x$ which corresponds to scaling the landmark coordinates by a scalar α. Every image contained within the training dataset is warped using piecewise affine transformation to register the high-resolution dictionary with the expected shape $\mathbf{z}_y$. The low-resolution dictionary is then constructed by simply down-sampling every image contained within the high-resolution dictionary by a scale factor α. In Fig. 8, we use LM-CSS method to super-resolve non-frontal views. Nevertheless, this method can be used to extend existing face hallucination techniques which can use the registered low- and high-resolution dictionaries to super-resolve the low-resolution test image $\mathbf{X}$ and synthesize the high-resolution face image $\widetilde{\mathbf{Y}}$. In this work we use the 21 facial landmark-points defined in [77] since it caters for both affine (rotation, translation and scaling) and more complicated 3-dimensional deformations from the frontal view.

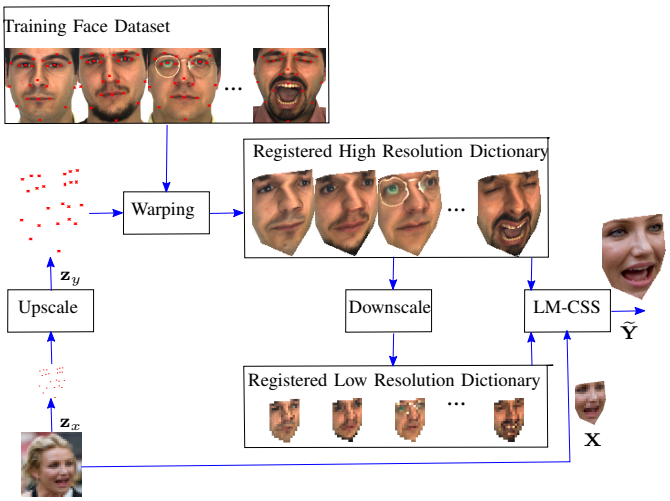


Fig. 8: Illustration of the LM-CSS in the Wild.

VI. Results

The proposed system is compared to several popular and/or state-of-the-art methods. In subsection VI-A we investigate the effect that the number of support points k has on performance for LM-CSS (Please note that the results shown in Section III are computed using the NE method [61]). The performance in terms of face recognition and quality are evaluated in sections VI-B and VI-C respectively. In all these experiments, the dictionary used to learn the up-scaling projection matrix for each patch consisted of a composite dataset which includes images from both Color Feret [63] and Multi-Pie [64] datasets, where only frontal facial images were considered. One image per subject was randomly selected, resulting in a dictionary of $m = 1203$ of facial images.

The gallery consisted of another composite dataset which combined frontal facial images from the FRGC-V2 (controlled environment) [78] and MEDS datasets [79]. One unique image per subject was randomly selected, providing a gallery of 889 facial images. The probe images were taken from the FRGC-V2 dataset (uncontrolled environment), where two images per subject were included, resulting in 930 probe images. Another gallery and probe set was considered from the CAS-PEAL dataset [], where the 1040 faces with neutral expression are used as gallery while 900 images with different expressions were randomly selected as probe images. All probe images are frontal faces, however various poses and illuminations were considered. All the images were registered using affine transformation computed on landmark points of the eyes and mouth centres, such that the distance between the eyes $d_y = 40$. The probe and low-resolution dictionary images were down-sampled to the desired scale α using MATLAB's $\mathrm{imresize}$ function.

The experiments in subsection VI-D were conducted to evaluate the performance of the proposed method in the wild. In this experiment we used the IJB-A dataset [80] which contains face images with a wide range of pose and orientation variations. In these experiments the AR dataset was used as a Training face dataset since it contains the 21 landmark points for each of the 886 subjects. Unless stated otherwise, all patch based methods are configured such that the number of pixels in a low-resolution patch $n = 25$, the low-resolution overlap $\gamma_x = 2$, and the patches are stitched by averaging overlapping regions. All the methods apply the super-resolution algorithm on the luminance component of the $\mathrm{YC}_b\mathrm{C}_r$ color model, while the chrominance components were up-scaled using bi-cubic interpolation. All simulations were run using a machine with Intel (R) Core (TM) i7-3687U CPU at 2.10GHz running Windows 64-bit Operating system.

A. Parameter Selection for LM-CSS

The proposed method has four parameters that need to be tuned, namely the regularization parameters $\lambda^{\{0\}}$, $\lambda^{\{1\}}$ and λ_s and the support size k. The regularization parameters $\lambda^{\{0\}}$ and $\lambda^{\{1\}}$ adopted by Multivariate Ridge Regression can be easily set to a very small value since its purpose is to perturb the linear-dependent vectors within a matrix to avoid singular values. In all experiments, these parameters were set to 10^{-6}. Similarly, the BPDN's regularization parameter λ_s which controls the sparsity of the solution was set to 0.01, since it provided satisfactory performance on the AR dataset.

Fig. 9 shows the average PSNR and Rank-1 recognition using the LBP face recognizer [65] on all 930 probe images using the FRGC dataset. From these results it can be observed that PSNR increases as the support size is increased, until $k = 150$ where it starts decreasing (or stays in steady state). This result confirms the observations obtained in section III where a different set of probe images was used. On the other hand, the best rank-1 recognition is attained at $k = 50$, and the recognition starts decreasing at larger values of k. Again, this confirms the results in section III where it was observed that more texture consistent facial images are obtained when using a smaller support size. We emphasize here that the results in section III were computed using the neighbour embedding [61] while the results presented in this subsection are computed for the proposed LM-CSS method.

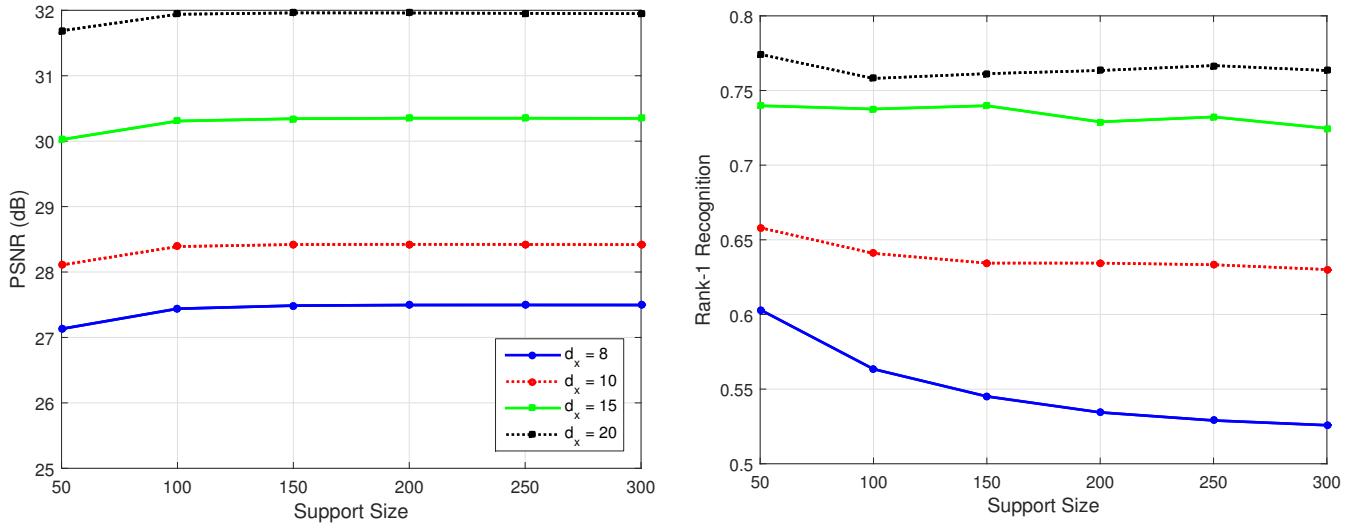
B. Recognition Analysis

Three face recognition methods were adopted in this experiment, namely the LBP face recognition [65] method (which was found to provide state-of-the-art performance on the single image per subject problem in [67]), the Gabor face recognizer [68][5] and the VGG-Face Convolutional Neural Network (CNN) descriptor [69][6]. The Gabor Face-recognizer method performs classification in the Principal Component Analysis (PCA) subspace, where the PCA basis were trained off-line on the AR dataset. The proposed method was compared with Bi-Cubic Interpolation and seven face hallucination methods, namely Eigen-transformation [14], Neighbour Embedding [61], the method of Yang *et. al*. (ScSR) [76], Position-Patch [33], Sparse Position-Patch, [34], Eigen-Patches [22] and LINE [47]. The LINE method, which up to the knowledge of the authors is the only method that tries to exploit the structure of the high-resolution manifold to bridge the low- and high-resolution manifolds, represents the current state-of-the-art in face hallucination. We also compare our method followed by an off-the-shelf face recognition system against the specialized very-low-resolution face recognition system DFD described in [58]. These methods were configured using the same patch size and overlap as indicated above and configured using the optimal parameters provided in their respective papers. The methods were implemented in MATLAB, where the code for [47], [58], [76] were provided by the authors. In this experiment we show results for LM-CSS with $k = 50$ and $k = 150$, where the latter corresponds to the neighbourhood size adopted by LINE.

The recognition performance is summarized in Tablea I and II on the FRGC and CAS-PEAL datasets respectively. In these tables we adopt the Area Under the ROC curve (AUC) as a scalar-valued measure of accuracy for unsupervised learning [81] together with the rank-1 recognition. The VGG-Face CNN face recognition was found to be particularly fragile at the very low-resolutions considered here, where it only achieves a rank-1 recognition up to 19% on bi-cubic interpolated images at a magnification factor $\alpha = 2$. It can also be noticed that face recognition can benefit from the texture detail recovered using face hallucination where rank-1 recognition rate gains between 20% and 60% were obtained over bi-cubic interpolated facial images when using the LBP face recognizer on very-low-resolution images. It can also be seen that the cross-resolution face recognition system DFD [58] performed well at low-resolutions such as $d_x = 20$, but its performance significantly drops at lower-resolutions where LM-CSS provided rank-1 recognition gains between and 34% and 60%. These results also show that our proposed method is

[5]The code was provided by the authors in http://www.mathworks.com/matlabcentral/fileexchange/35106-the-phd-face-recognition-toolbox.

[6]The pre-trained network was provided by the authors in http://www.robots.ox.ac.uk/~vgg/software/vgg_face/



(a) Quality Analysis

(b) Recognition Analysis

Fig. 9: Analysing the effect of the number of support vectors on the performance of the proposed LM-CSS hallucination method on the FRGC dataset.

most of the time superior to all the other methods considered in this experiment in terms of both rank-1 recognition and AUC metric which involves a total of 1830 probe images from two different datasets. More precisely, rank-1 recognition rate gains of between 1% and 2% were achieved over LINE and between 2% and 8% over Eigen-Patches when using the LBP face recognizer.

Figure 10 shows the Cumulative Matching Score Curve (CMC) which measures the recognition at different ranks for the FRGC dataset. For clarity, only the methods that showed competitive rank-1 recognition rates were included, and compared to the oracle where the probe images were not down-sampled (*i.e.* $d_x = d_y$). Bi-cubic interpolation was included to show that state-of-the-art face recognition methods benefit from the texture details recovered by patch-based hallucination methods, achieving significant higher recognition rates at all ranks. It can also be noted that the proposed method outperforms the other hallucination methods, especially at lower resolutions and lower ranks.

C. Quality Analysis

Tables III and IV show the quality analysis measured in terms of PSNR and the Structural Similarity Metric (SSIM) [66] on the FRGC and CAS-PEAL datasets respectively. In this experiment we compare our LM-CSS method (configured using $k = 50$ and $k = 150$). These results show that the proposed method outperforms all other schemes on the FRGC dataset and is competitive to Eigen-Patches and superior to all other schemes on the CAS-PEAL dataset. It is important to notice that while LINE provided the most competitive results in terms of recognition, its performance in terms of PSNR is relatively low, where LM-CSS achieves PSNR gains of 1 dB at different magnification factors on the CAS-PEAL dataset, and we achieve better PSNR even when using a smaller number of support *i.e.* $k = 50$. This can be attributed to the fact that since LINE employs a neighbourhood size 150 it manages to recover important texture detail, but the resulting face image is noisy. On the other hand, Eigen-Patches adopts all atoms in the dictionary and achieves good PSNR performance at the expense of getting images which are more blurred and lacks important texture details. This can but confirmed by its poor performance in terms of recognition. Subjective results in the following subsection consolidate these remarks. It is also important to mention that the dictionary which is used to restore facial images from both datasets is made up of mixed ethnicities. Nevertheless, this dictionary provides good performance on the CAS-PEAL dataset which only contains faces of Asian people.

D. Face Hallucination in the Wild

In order to assess the performance of the proposed method in the wild, we conduct experiments using 104 images from the IJB-A dataset. In this experiment, we compare our proposed LM-CSS method with LINE [47] and Eigen-Patches [22] which provided the most competitive performance in terms of recognition and PSNR respectively. These methods are not adequate to perform face hallucination in the wild since they are inherently designed to super-resolve frontal faces. Nevertheless, all these schemes were extended to perform face hallucination in the wild using the method described in Section V. All the images in the AR dataset are used as a dictionary here since they contain all the 21 landmark points that are required to register the training images to the non-frontal test images. Every image was scaled so that the high-resolution faces have an inter-eye distance $d_y = 40$ and the low-resolution face images have an inter-eye distance $d_x = 8$. We also compare these methods with two state-of-the-art generic super-resolution methods, namely the Convolutional Neural Network based scheme (SRCNN)m [9] and Non-local

TABLE I: Summary of the Rank-1 recognition results and Area Under Curve (AUC) metric using three different face recognition algorithms on the FRGC dataset.

Hall^n Method	**Rec^n Method**	**Resolution d_x**							
		8		**10**		**15**		**20**	
		rank-1	**AUC**	**rank-1**	**AUC**	**rank-1**	**AUC**	**rank-1**	**AUC**
Bi-Cubic	Gabor	0.0000	0.6985	0.0000	0.7823	0.0344	0.8829	0.5215	0.9181
	LBP	0.3065	0.9380	0.5032	0.9598	0.6065	0.9708	0.7054	0.9792
	DeepFaces	0.0000	0.5296	0.0000	0.5337	0.0258	0.6223	0.1903	0.7157
Eigentransformation [14]	Gabor	0.0591	0.7852	0.1097	0.8359	0.3312	0.8841	0.5183	0.9098
	LBP	0.2559	0.9390	0.4516	0.9554	0.5624	0.9633	0.6495	0.9688
	DeepFaces	0.0151	0.5794	0.0194	0.5954	0.0398	0.6187	0.1226	0.6612
Neighbour Embedding [61]	Gabor	0.2323	0.8624	0.4710	0.8968	0.6172	0.9182	0.6409	0.9272
	LBP	0.5548	0.9635	0.6398	0.9712	0.7215	0.9795	0.7559	0.9830
	DeepFaces	0.0151	0.5940	0.0602	0.6290	0.2086	0.6970	0.4075	0.7562
ScSR [76]	Gabor	0.0000	0.7618	0.0065	0.8392	0.4237	0.9048	0.6215	0.9250
	LBP	0.4237	0.9486	0.6000	0.9650	0.6860	0.9765	0.7559	0.9823
	DeepFaces	0.0000	0.5399	0.0000	0.5455	0.0441	0.6364	0.2376	0.7200
Sparse Position-Patches [34]	Gabor	0.2333	0.8632	0.4645	0.8969	0.6118	0.9152	0.6398	0.9254
	LBP	0.5677	0.9649	0.6441	0.9721	0.7247	0.9803	0.7570	0.9830
	DeepFaces	0.0161	0.5870	0.0419	0.6198	0.1624	0.6880	0.3796	0.7553
Position-Patches [33]	Gabor	0.1108	0.8354	0.2849	0.8814	0.5774	0.9154	0.6419	0.9281
	LBP	0.4699	0.9588	0.5849	0.9675	0.6849	0.9782	0.7312	0.9812
	DeepFaces	0.0161	0.5914	0.0398	0.6201	0.1785	0.6878	0.3559	0.7408
Eigen-Patches [22]	Gabor	0.1613	0.8517	0.3849	0.8934	0.6065	0.9172	0.6387	0.9283
	LBP	0.5226	0.9625	0.6215	0.9704	0.7237	0.9800	0.7602	0.9830
	DeepFaces	0.0129	0.5927	0.0452	0.6226	0.1882	0.6900	0.3516	0.7457
LINE [47]	Gabor	0.3118	0.8696	0.5011	0.8986	0.6118	0.9168	0.6409	0.9252
	LBP	0.5925	0.9647	0.6559	0.9714	0.7323	**0.9804**	0.7677	**0.9833**
	DeepFaces	0.0312	0.6036	0.0710	0.6385	0.2161	0.7050	0.4172	0.7630
DFD [58]		0.0108	0.7387	0.1570	0.8684	0.5978	0.9559	0.7677	0.9772
Proposed ($k = 50$)	Gabor	0.2753	0.8803	0.5000	0.9036	0.6183	0.9202	0.6452	0.9281
	LBP	**0.6032**	**0.9658**	**0.6581**	**0.9722**	**0.7398**	0.9798	**0.7742**	**0.9833**
	DeepFaces	0.0172	0.5874	0.0484	0.6293	0.1914	0.7015	0.4022	0.7609
Proposed ($k = 150$)	Gabor	0.1978	0.8701	0.4495	0.8996	0.6140	0.9201	0.6409	0.9292
	LBP	0.5452	0.9644	0.6344	0.9710	**0.7398**	0.9801	0.7602	0.9831
	DeepFaces	0.0151	0.5890	0.0527	0.6291	0.2108	0.7011	0.3828	0.7578

Centralized Sparse Representations (NCSR) [7]. From the results in Fig. 11 one can see that the generic super-resolution methods manage to improve the quality of the original LR image, but the resulting faces are generally blurred since they do not exploit the facial structure. On the other hand, the facial images reconstructed using Eigen-Patches (EP) are blurred and lacks texture detail which is important for recognition, which confirms its poor performance in terms of recognition. On the other hand, the results provided by LINE contain sever structural noise, which confirms our hypothesis that it is not guaranteed to converge to an optimal solution, thus confirming the poor PSNR performance in Section VI-C. On the other hand, the LM-CSS method manages to reconstruct to get facial images which are closer to the ground truth with more texture detail compared to SRCNN, NCSR and Eigen-Patches, and less structural noise when compared to LINE.

The input low-resolution face images of the above experiments are formed by smoothing and down-sampling the original high-resolution face image to be able to compare it with respect to the ground-truth. This does not represent the real relationship between the unknown high-resolution and the available low-resolution face image in the real world [82]. In order to further assess the effectiveness of the proposed method, we conduct experiments on some real low-resolution face image from the IJB-A dataset and compare the actual low-resolution face image with the proposed LM-CSS solution in Fig. 12. Apart from the distortions caused by blurring and down-sampling, the test images contain compression artefacts which are not catered by our proposed method. Nevertheless, it can be seen that even though the proposed LM-CSS in the wild does not cater for compression artefacts, it manages to increase the texture detail on the facial region. Moreover, in our previous work we demonstrated that dictionary based super-resolution methods, such as LM-CSS, can be made robust to compression by exploiting the syntax of the compressed image/video [83].

E. Complexity Analysis

The complexity in terms of the average time taken to synthesize a high-resolution image from a low-resolution image in seconds is summarized in Table V. These results show that the proposed method is significantly less computationally intensive than Eigen-Patches but more complex than the other methods, including LINE. While complexity is not the prime aim of this work, the performance of the proposed scheme can be significantly improved using more efficient l_1-minimization algorithms to solve the problem in *Layer 1* as mentioned in [76], since this is the most computationally intensive part in our method.

VII. Conclusion

In this paper, we propose a new approach which can be used to synthesize a high-resolution facial image from a low-resolution test image. The proposed method first derives a smooth approximation which is close to the ground-truth in

TABLE II: Summary of the Rank-1 recognition results and Area Under Curve (AUC) metric using two different face recognition algorithms (on the CAS-PEAL dataset.

Hall^n Method	**Rec^n Method**	**Resolution d_x**							
		8		**10**		**15**		**20**	
		rank-1	**AUC**	**rank-1**	**AUC**	**rank-1**	**AUC**	**rank-1**	**AUC**
Bi-Cubic	Gabor	0.0156	0.7928	0.0300	0.8470	0.1811	0.8729	0.4211	0.8775
	LBP	0.2111	0.8945	0.4656	0.9385	0.5033	0.9420	0.7622	0.9715
	DeepFaces	0.0033	0.5437	0.0033	0.5527	0.0533	0.6353	0.1422	0.6857
Eigentransformation [14]	Gabor	0.0400	0.7542	0.0956	0.8078	0.2467	0.8489	0.4167	0.8644
	LBP	0.1500	0.8410	0.3367	0.8897	0.3511	0.8960	0.5489	0.9218
	DeepFaces	0.0144	0.5723	0.0156	0.5760	0.0300	0.5897	0.0356	0.6082
Neighbour Embedding [61]	Gabor	0.1356	0.8264	0.3667	0.8767	0.5700	0.8992	0.6433	0.8905
	LBP	0.5567	0.9213	0.6622	0.9457	0.8067	0.9745	0.8556	0.9796
	DeepFaces	0.0156	0.5846	0.0267	0.6052	0.0878	0.6670	0.1400	0.7069
ScSR [76]	Gabor	0.0356	0.8298	0.0967	0.8617	0.3644	0.8728	0.5456	0.8773
	LBP	0.2800	0.9061	0.5656	0.9536	0.6056	0.9538	0.8122	0.9753
	DeepFaces	0.0044	0.5517	0.0089	0.5648	0.0567	0.6217	0.1433	0.6809
Sparse Position-Patches [34]	Gabor	0.1467	0.8268	0.3100	0.8670	0.5356	0.8795	0.6100	0.8802
	LBP	0.5400	0.9282	0.6633	0.9525	0.8211	0.9668	0.8667	0.9709
	DeepFaces	0.0178	0.5818	0.0244	0.5996	0.0922	0.6584	0.1644	0.7013
Position-Patches [33]	Gabor	0.0933	0.8355	0.2411	0.8798	0.5178	0.8916	0.6156	0.8880
	LBP	0.4711	0.9114	0.5889	0.9452	0.7800	0.9662	0.8356	0.9701
	DeepFaces	0.0200	0.5992	0.0333	0.6196	0.1022	0.6821	0.1544	0.7085
Eigen-Patches [22]	Gabor	0.1422	0.8451	0.3322	0.8829	0.5600	0.8896	0.6300	0.8867
	LBP	0.5400	0.9192	0.6567	0.9497	0.8222	0.9726	0.8611	0.9708
	DeepFaces	0.0167	0.5950	0.0422	0.6148	0.1011	0.6675	0.1678	0.7078
LINE [47]	Gabor	0.1700	0.8303	0.4056	0.8722	0.5822	0.8920	0.6322	0.8849
	LBP	**0.5822**	0.9220	0.6833	0.9458	0.8167	0.9744	0.8633	0.9797
	DeepFaces	0.0144	0.5888	0.0267	0.6065	0.1033	0.6672	0.1489	0.7057
DFD [58]		0.0200	0.6865	0.1022	0.8299	0.4778	0.9331	0.8467	0.9716
Proposed ($k = 50$)	Gabor	0.1811	0.8427	0.3633	0.8733	0.5656	0.8888	0.6311	0.8896
	LBP	0.5622	0.9282	**0.6922**	0.9538	0.8256	0.9756	0.8689	0.9806
	DeepFaces	0.0178	0.5851	0.0289	0.6015	0.0989	0.6597	0.1622	0.7067
Proposed ($k = 150$)	Gabor	0.1544	0.8516	0.3344	0.8789	0.5644	0.8900	0.6244	0.8904
	LBP	0.5344	**0.9385**	0.6789	**0.9611**	**0.8322**	**0.9783**	**0.8756**	**0.9820**
	DeepFaces	0.0189	0.5969	0.0356	0.6117	0.1122	0.6726	0.1722	0.7086

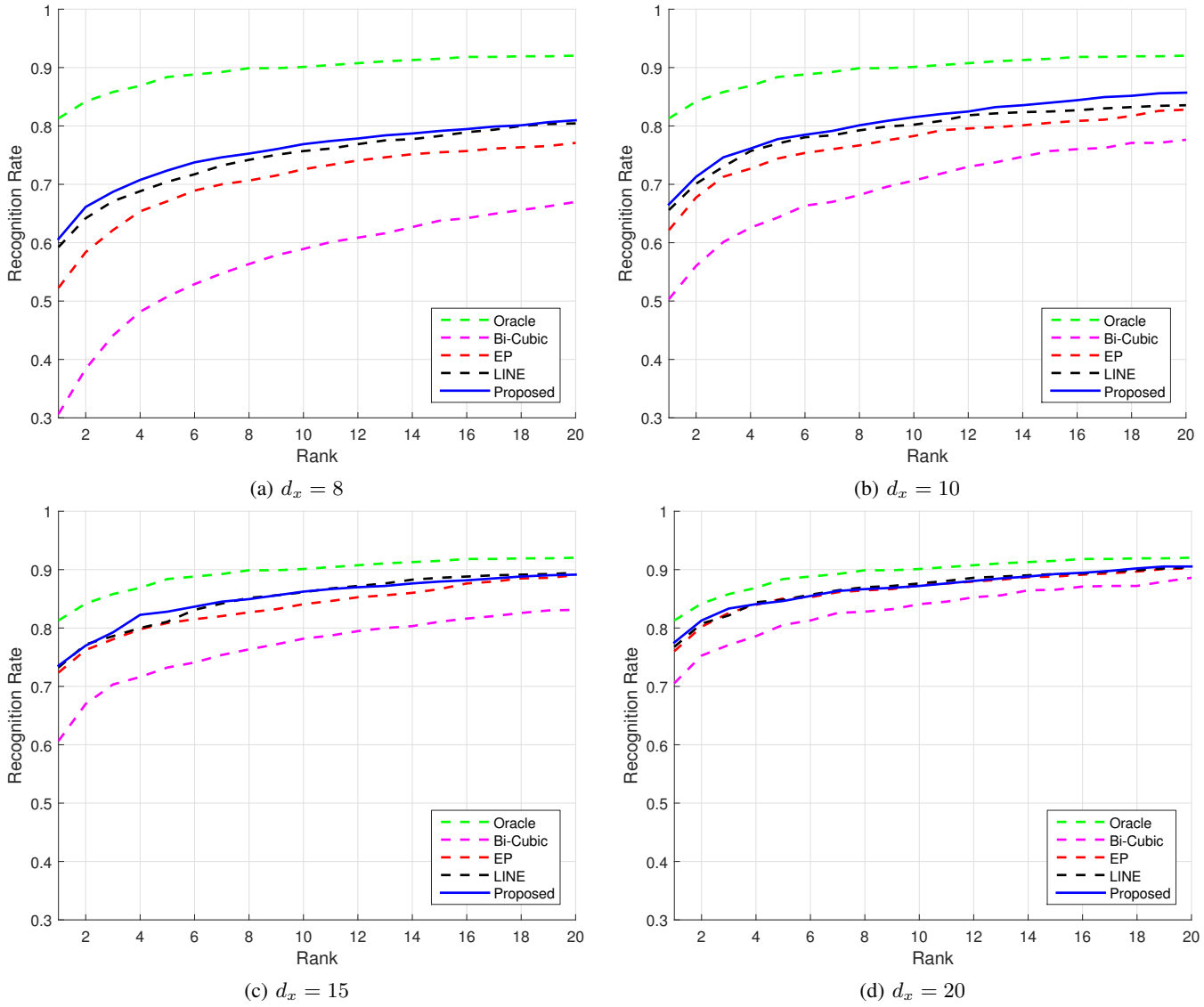
TABLE III: Summary of the Quality Analysis results using the PSNR and SSIM quality metrics on the FRGC dataset.

Hall^n Method	**Resolution d_x**							
	8		**10**		**15**		**20**	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
Bi-Cubic	24.0292	0.6224	26.2024	0.7338	25.2804	0.7094	28.6663	0.8531
Eigentransformation [14]	24.3958	0.6496	26.8645	0.7504	24.9374	0.6724	27.7883	0.7892
Neighbour Embedding [61]	26.9987	0.7533	27.9560	0.7973	29.9892	0.8714	31.6301	0.9122
ScSR [76]	24.1088	0.6316	26.6015	0.7600	24.9749	0.7067	28.7794	0.8639
Position-Patches [33]	27.3044	0.7731	28.2906	0.8145	30.1887	0.8785	31.7192	0.9143
Sparse Position-Patches [34]	27.2500	0.7666	28.2219	0.8100	30.1290	0.8767	31.7162	0.9146
Eigen-Patches [22]	27.3918	0.7778	28.3847	**0.8196**	30.3118	0.8842	31.8986	0.9203
LINE [47]	27.0927	0.7591	28.0253	0.8009	30.0471	0.8727	31.6970	0.9131
Proposed ($k = 50$)	27.1307	0.7679	28.1078	0.8093	30.0240	0.8761	31.6875	0.9139
Proposed ($k = 150$)	**27.4866**	**0.7802**	**28.4200**	0.8009	**30.3431**	**0.8845**	**31.9610**	**0.9209**

TABLE IV: Summary of the Quality Analysis results using the PSNR and SSIM quality metrics on the CAS-PEAL dataset.

Hall^n Method	**Resolution d_x**							
	8		**10**		**15**		**20**	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
Bi-Cubic	22.1115	0.6517	25.0259	0.7831	23.3753	0.7478	27.9440	0.9016
Eigentransformation [14]	21.8434	0.6324	24.7969	0.7457	22.7048	0.6681	26.0714	0.7913
Neighbour Embedding [61]	25.1745	0.7545	26.7105	0.8166	29.7110	0.9078	32.3148	0.9488
ScSR [76]	21.9956	0.6589	25.4175	0.8029	22.9500	0.7406	27.9659	0.9078
Position-Patches [33]	25.9597	0.8036	27.5107	0.8577	30.3056	0.9250	32.7771	0.9569
Sparse Position-Patches [34]	25.6982	0.7825	27.2117	0.8418	30.1578	0.9192	32.6840	0.9546
Eigen-Patches [22]	26.0834	**0.8071**	**27.6536**	**0.8620**	30.6109	**0.9303**	33.1584	**0.9615**
LINE [47]	25.2454	0.7567	26.7365	0.8176	29.8449	0.9085	32.4446	0.9492
Proposed ($k = 50$)	25.5597	0.7704	27.0966	0.8318	30.0366	0.9125	32.6052	0.9521
Proposed ($k = 150$)	**26.0875**	0.8046	27.6003	0.8579	**30.6199**	0.9287	**33.2108**	0.9613

Euclidean space on the high-resolution manifold. Based on the assumption that the patches reside on a high-resolution manifold, we assume that the optimal support to represent the first approximation is good to reconstruct the ground-truth, which was shown to be valid in other research domains [70], [71]. The coupled sparse support is then used to model



(a) $d_x = 8$ (b) $d_x = 10$

(c) $d_x = 15$ (d) $d_x = 20$

Fig. 10: Cumulative Matching Score Curves (CMC) of face images hallucinated from different resolutions d_x.

TABLE V: Summary of the time taken (in seconds) to synthesize one image at different resolutions.

Hall^n Method	**Resolution d_x**			
	8	**10**	**15**	**20**
Eigentransformation [14]	2.84	2.69	2.74	2.75
Neighbour Embedding [61]	0.25	0.33	0.73	1.24
Position-Patches [33]	1.59	2.11	4.69	8.37
Sparse Position-Patches [34]	0.59	0.79	1.72	3.02
Eigen-Patches [22]	10.89	14.80	34.74	63.98
LINE [47]	2.23	2.16	2.61	2.83
Proposed	8.04	5.78	4.71	5.43

the up-scaling function for each patch using Multivariate Ridge Regression. The proposed method differs from existing methods since i) it models the up-scaling function for each patch rather than combining a number of high-resolution faces which makes it more acceptable by the forensics community, ii) the proposed method exploits the local structure of the high-resolution manifold to select the optimal support instead of exploiting the structure of the low-resolution manifold which is known to be distorted, iii) LM-CSS is non-iterative and is guaranteed to converge to an optimal solution and iv) we propose a method that can extend face hallucination methods to be used on non-frontal images.

Extensive simulations were conducted on frontal images from the FRGC and CAS-PEAL dataset, where a total of 1830 images were evaluated in terms of both recognition and quality. This makes the most extensive evaluation of face hallucination that can be found in literature. From these results it was concluded that images super-resolved using the proposed LM-CSS followed by LBP face recognition provides the best recognition performance. It was found to significantly outperform the cross-resolution face recognition system DFD [58] where rank-1 recognition gains between 34% and 60% were attained at very low-resolution. It was also shown to outperform other face hallucination schemes followed by LBP. The quality analysis shows that our LM-CSS method outperforms existing methods, most of the time. Moreover, subjective results show that apart from outperforming generic

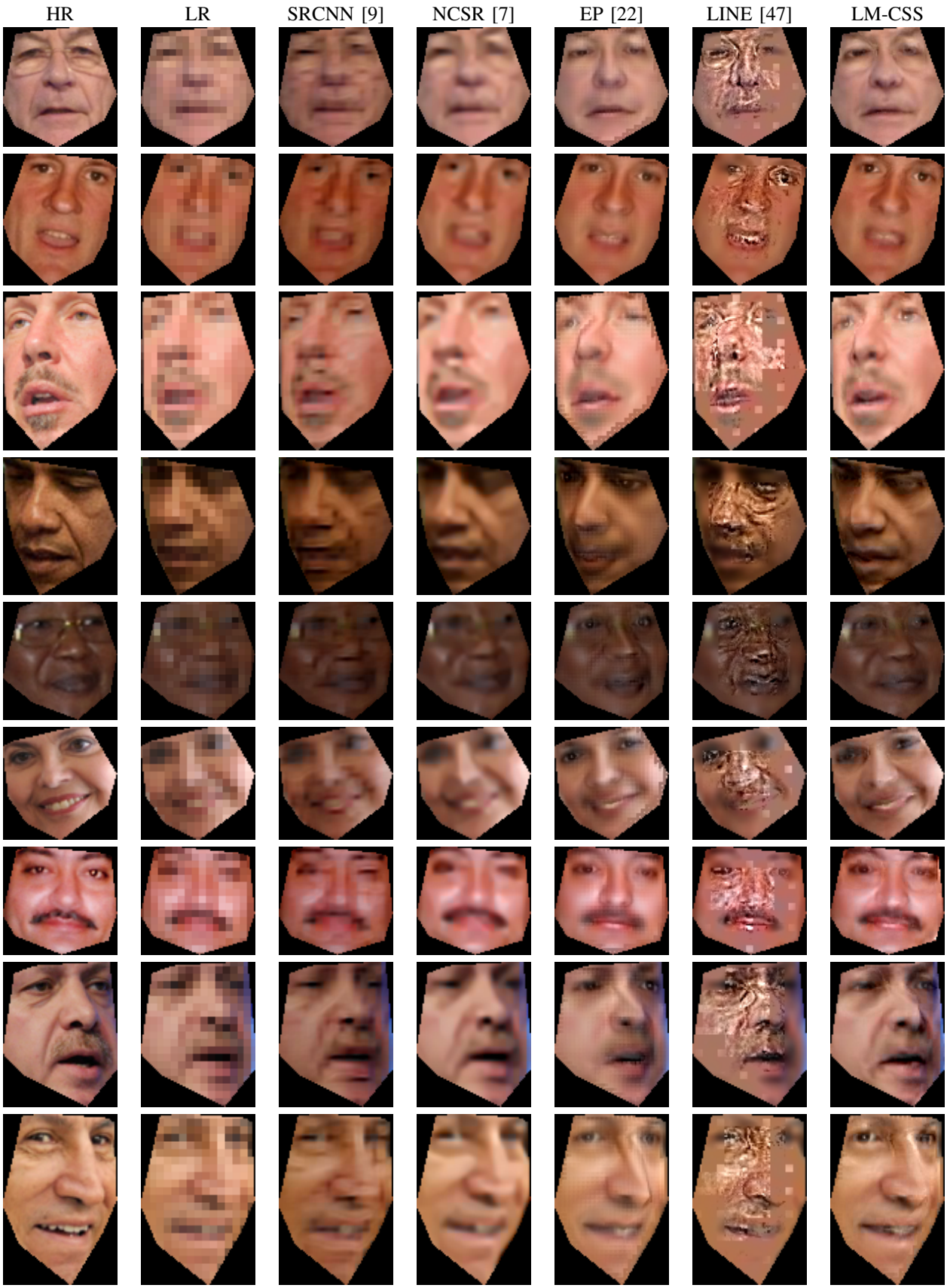


Fig. 11: Super-resolved face images in the wild. The HR image represents the original ground-truth and LR is the low-resolution image where $d_x = 8$ and the other images are restored with a magnification factor $\alpha = 5$.

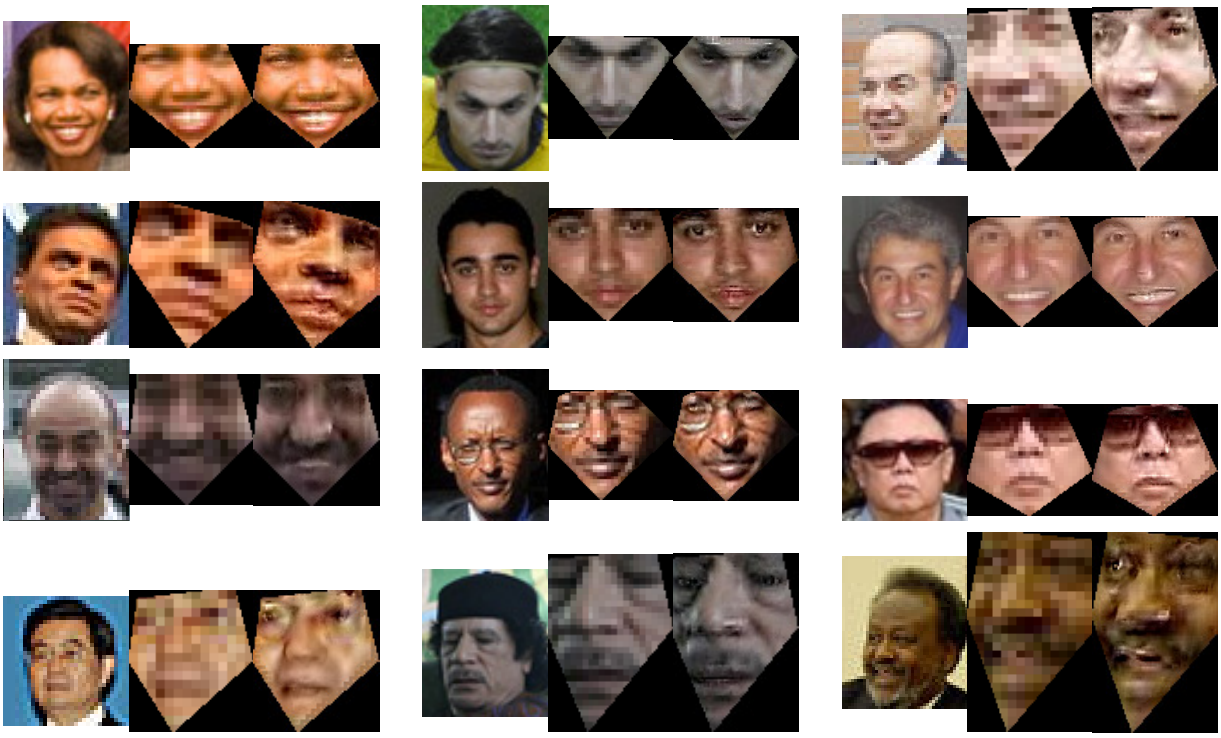


Fig. 12: Super-resolved results from real-world low-resolution faces taken from the IJB-A dataset. For each example, the original LR face is shown on the left, the facial region of the low resolution face image is shown at the centre and the synthesized high-resolution face image is shown at the right.

super-resolution in the wild, our LM-CSS manages to recover higher quality images than both LINE and Eigen-Patches, which showed the most competitive results.

Future work points us in the direction to implement face hallucination techniques which are able to hallucinate and enhance face images afflicted by different distortions such as compression, landmark-point misalignment, bad exposure and other distortions commonly found in CCTV images. The ability of current schemes (including the proposed method) are dependent on the dictionaries used, and therefore these schemes can be made more robust by building more robust dictionaries.

Acknowledgment

The authors would like to thank the authors in [47], [58], [68], [76] for providing the source code of their methods and Prof. David Donoho and his team at SparseLab for providing the BPDN l_1-minimization solver. The authors would also like to thank the reviewers who have helped us improve the quality of the paper.

References

[1] W. Zou and P. Yuen, "Very low resolution face recognition problem," *IEEE Trans. on Image Processing*, vol. 21, no. 1, pp. 327–340, Jan 2012.

[2] M. A. Sasse, "Not seeing the crime for the cameras?" *ACM Commun.*, vol. 53, no. 2, pp. 22–25, Feb. 2010.

[3] N. La Vigne, S. Lowry, J. Markman, and A. Dwyer, "Evaluating the use of public surveillance cameras for crime control and prevention," Urban Institute Justice Policy Center, Tech. Rep., 2011.

[4] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. of Computer Vision*, vol. 106, no. 1, pp. 9–30, 2014.

[5] M. Elad and A. Feuer, "Super-resolution reconstruction of image sequences," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 817–834, Sep 1999.

[6] S. Farsiu, M. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. on Image Processing*, vol. 13, no. 10, pp. 1327–1344, Oct 2004.

[7] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Trans. on Image Processing*, vol. 22, no. 4, pp. 1620–1630, Apr. 2013.

[8] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE Trans. on Image Processing*, vol. 23, no. 6, pp. 2569–2582, June 2014.

[9] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb 2016.

[10] Y. Hu, N. Wang, D. Tao, X. Gao, and X. Li, "Serf: A simple, effective, robust, and fast image super-resolver from cascaded linear regression," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4091–4102, Sept 2016.

[11] S. Baker and T. Kanade, "Hallucinating faces," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2000, pp. 83–88.

[12] Y. Hu, K.-M. Lam, G. Qiu, and T. Shen, "From local pixel structure to global image super-resolution: A new face hallucination framework," *IEEE Trans. on Image Processing*, vol. 20, no. 2, pp. 433–445, Feb 2011.

[13] Y. Li, C. Cai, G. Qiu, and K.-M. Lam, "Face hallucination based on sparse local-pixel structure," *Pattern Recognition*, vol. 47, no. 3, pp. 1261 – 1270, 2014.

[14] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. on Systems, Man, and Cybernetics, Part C*, vol. 35, no. 3, pp. 425–434, Aug 2005.

[15] J.-S. Park and S.-W. Lee, "An example-based face hallucination method for single-frame, low-resolution facial images," *IEEE Trans. on Image Processing*, vol. 17, no. 10, pp. 1806–1816, Oct 2008.

[16] C. Liu, H.-Y. Shum, and C.-S. Zhang, "A two-step approach to halluci-

nating faces: global parametric model and local nonparametric model," in *Proc. on IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2001.

[17] Y. Li and X. Lin, "An improved two-step approach to hallucinating faces," in *Image and Graphics (ICIG'04), Third International Conference on*, Dec 2004, pp. 298–301.

[18] C. Liu, H.-Y. Shum, and W. Freeman, "Face hallucination: Theory and practice," *Int. J. of Computer Vision*, vol. 75, no. 1, pp. 115–134, 2007.

[19] A. Chakrabarti, A. Rajagopalan, and R. Chellappa, "Super-resolution of face images using kernel pca-based prior," *IEEE Trans. on Multimedia*, vol. 9, no. 4, pp. 888–892, June 2007.

[20] Y. Zhuang, J. Zhang, and F. Wu, "Hallucinating faces: {LPH} super-resolution and neighbor reconstruction for residue compensation," *Pattern Recognition*, vol. 40, no. 11, pp. 3178 – 3194, 2007.

[21] J. Yang, H. Tang, Y. Ma, and T. Huang, "Face hallucination via sparse coding," in *Proc. of IEEE Conf. on Image Processing*, Oct 2008, pp. 1264–1267.

[22] H.-Y. Chen and S.-Y. Chien, "Eigen-patch: Position-patch based face hallucination using eigen transformation," in *Proc. on IEEE Int. Conf. on Multimedia and Expo*, July 2014, pp. 1–6.

[23] X. Zhang, S. Peng, and J. Jiang, "An adaptive learning method for face hallucination using locality preserving projections," in *Proc. on IEEE Conf. on Automatic Face Gesture Recognition*, Sept 2008, pp. 1–8.

[24] B. Kumar and R. Aravind, "Face hallucination using olpp and kernel ridge regression," in *Proc. on IEEE Conf. on Image Processing*, Oct 2008, pp. 353–356.

[25] W. Liu, D. Lin, and X. Tang, "Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation," in *Proc. on IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, June 2005, pp. 478–484 vol. 2.

[26] N. Wang, J. Li, D. Tao, X. Li, and X. Gao, "Heterogeneous image transformation," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 77 – 84, 2013, extracting Semantics from Multi-Spectrum Video.

[27] T. Lu, R. Hu, Z. Han, J. Jiang, and Y. Xia, "Robust super-resolution for face images via principle component sparse representation and least squares regression," in *Proc. on IEEE Int. Symp. on Circuits and Systems*, May 2013, pp. 1199–1202.

[28] K. Su, Q. Tian, Q. Xue, N. Sebe, and J. Ma, "Neighborhood issue in single-frame image super-resolution," in *Proc. on IEEE Int. Conf. on Multimedia and Expo*, July 2005, pp. 4 – 7.

[29] W. Liu, D. Lin, and X. Tang, "Neighbor combination and transformation for hallucinating faces," in *Proc. on IEEE Int. Conf. on Multimedia and Expo*, July 2005.

[30] W. Zhang and W.-K. Cham, "Hallucinating face in the dct domain," *IEEE Trans. on Image Processing*, vol. 20, no. 10, pp. 2769–2779, Oct 2011.

[31] X. Du, F. Jiang, and D. Zhao, "Multi-scale face hallucination based on frequency bands analysis," in *Proc. on IEEE Conf. on Visual Communications and Image Processing*, Nov 2013, pp. 1–6.

[32] S. W. Park and M. Savvides, "Breaking the limitation of manifold analysis for super-resolution of facial images," in *Proc. on IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, April 2007, pp. I–573–I–576.

[33] X. Ma, J. Zhang, and C. Qi, "Position-based face hallucination method," in *Proc. on IEEE Int. Conf. on Multimedia and Expo*, June 2009, pp. 290–293.

[34] C. Jung, L. Jiao, B. Liu, and M. Gong, "Position-patch based face hallucination using convex optimization," *IEEE Signal Processing Letters*, vol. 18, no. 6, pp. 367–370, June 2011.

[35] J. Jiang, R. Hu, Z. Han, T. Lu, and K. Huang, "Position-patch based face hallucination via locality-constrained representation," in *Proc. on IEEE Int. Conf. on Multimedia and Expo*, July 2012, pp. 212–217.

[36] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Trans. on Multimedia*, vol. 16, no. 5, pp. 1268–1281, Aug 2014.

[37] H. Li, L. Xu, and G. Liu, "Face hallucination via similarity constraints," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 19–22, Jan 2013.

[38] S. Qu, R. Hu, S. Chen, Z. Wang, J. Jiang, and C. Yang, "Face hallucination via cauchy regularized sparse representation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 1216–1220.

[39] T. Lu, Z. Xiong, Y. Wan, and W. Yang, "Face hallucination via locality-constrained low-rank representation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 1746–1750.

[40] B. Li, H. Chang, S. Shan, and X. Chen, "Locality preserving constraints for super-resolution with neighbor embedding," in *Proc. on IEEE Int. Conf. on Image Processing*, Nov 2009, pp. 1189–1192.

[41] ——, "Aligning coupled manifolds for face hallucination," *IEEE Signal Processing Letters*, vol. 16, no. 11, pp. 957–960, Nov 2009.

[42] Y. Hao and C. Qi, "Face hallucination based on modified neighbor embedding and global smoothness constraint," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1187–1191, Oct 2014.

[43] W. Liu, D. Lin, and X. Tang, "Face hallucination through dual associative learning," in *Proc. on IEEE Conf. on Image Processing*, vol. 1, Sept 2005, pp. I–873–6.

[44] H. Huang, H. He, X. Fan, and J. Zhang, "Super-resolution of human face image using canonical correlation analysis," *Pattern Recognition*, vol. 43, no. 7, pp. 2532 – 2543, 2010.

[45] J. Jiang, R. Hu, Z. Han, Z. Wang, T. Lu, and J. Chen, "Locality-constraint iterative neighbor embedding for face hallucination," in *Proc. on IEEE Int. Conf. on Multimedia and Expo*, July 2013, pp. 1–6.

[46] S. Qu, R. Hu, S. Chen, J. Jiang, Z. Wang, and J. Chen, "Face hallucination via re-identified k-nearest neighbors embedding," in *Proc. on IEEE Int. Conf. on Multimedia and Expo*, July 2014, pp. 1–6.

[47] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multi-layer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Trans. on Image Processing*, vol. 23, no. 10, pp. 4220–4231, Oct 2014.

[48] S. Kolouri and G. Rohde, "Transport-based single frame super resolution of very low resolution face images," in *Proc. on IEEE Conf. on Computer Vision and Pattern Recognition*, June 2015, pp. 4876–4884.

[49] C.-T. Tu and J.-R. Luo, "Robust face hallucination using ensemble of feature-based regression functions and classifiers," *Image and Vision Computing*, vol. 44, pp. 59 – 72, 2015.

[50] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15. AAAI Press, 2015, pp. 3871–3877. [Online]. Available: http://dl.acm.org/citation.cfm?id=2888116.2888253

[51] P. Hennings-Yeomans, S. Baker, and B. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Proc. on IEEE Conf. on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[52] H. Bhatt, R. Singh, M. Vatsa, and N. Ratha, "Improving cross-resolution face matching using ensemble-based co-transfer learning," *IEEE Trans. on Image Processing*, vol. 23, no. 12, pp. 5654–5669, Dec 2014.

[53] M. Jian and K.-M. Lam, "Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 25, no. 11, pp. 1761–1772, Nov 2015.

[54] B. Li, H. Chang, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 20–23, Jan 2010.

[55] C. Zhou, Z. Zhang, D. Yi, Z. Lei, and S. Li, "Low-resolution face recognition via simultaneous discriminant analysis," in *Proc. on Joint Conf. on Biometrics (IJCB)*, Oct 2011, pp. 1–6.

[56] S. Siena, V. Boddeti, and B. Vijaya Kumar, "Coupled marginal fisher analysis for low-resolution face recognition," in *Proc. on European Conf. on Computer Vision*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7584, pp. 240–249.

[57] S. Siena, V. Boddeti, and B. Kumar, "Maximum-margin coupled mappings for cross-domain matching," in *IEEE Int. Conf. on Biometrics: Theory, Applications and Systems*, Sept 2013, pp. 1–8.

[58] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 289–302, 2014.

[59] N. Wang, X. Gao, J. Li, B. Song, and Z. Li, "Evaluation on synthesized face sketches," *Neurocomputing*, vol. 214, pp. 991 – 1000, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231216307603

[60] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognition*, vol. 39, no. 9, pp. 1725 – 1745, 2006.

[61] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. on IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, June 2004, pp. I–I.

[62] A. Martinez and R. Benavente, "The ar face database," Robot Vision Lab, Purdue University, Tech. Rep. 5, Apr. 1998.

[63] Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, Apr. 1998.

[64] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vision Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.

[65] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[66] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[67] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, Feb 2012.

[68] V. Štruc and N. Pavešić, "The complete gabor-fisher classifier for robust face recognition," *EURASIP J. Adv. Signal Process*, vol. 2010, pp. 31:1–31:13, Feb. 2010.

[69] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[70] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. on IEEE Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 3360–3367.

[71] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2009, pp. 2223–2231.

[72] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000. [Online]. Available: http://science.sciencemag.org/content/290/5500/2323

[73] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.

[74] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. on Information Theory*, vol. 52, no. 1, pp. 6–18, Jan 2006.

[75] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. on Conf. on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '01. New York, NY, USA: ACM, 2001, pp. 341–346.

[76] A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Fast l1-minimization algorithms and an application in robust face recognition: A review," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2010-13, Feb 2010.

[77] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *IEEE Int. Conf. on Computer Vision*, Nov 2011, pp. 2144–2151.

[78] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. on IEEE Conf. on Computer Vision and Pattern Recognition*, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 947–954.

[79] A. P. Founds, N. Orlans, G. Whiddon, and C. Watson, "Nist special database 32 multiple encounter dataset ii (meda-ii)," National Institute of Standards and Technology, Tech. Rep., 2011.

[80] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *Proc. on IEEE Conf. on Computer Vision and Pattern Recognition*, June 2015, pp. 1931–1939.

[81] G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," University of Massachusetts, Tech. Rep., 2014.

[82] P. P. Gajjar and M. V. Joshi, "New learning based super-resolution: Use of dwt and igmrf prior," *IEEE Trans. on Image Processing*, vol. 19, no. 5, pp. 1201–1213, May 2010.

[83] R. Farrugia and C. Guillemot, "Robust face hallucination using quantization-adaptive dictionaries," in *IEEE Int. Conf. on Image Processing*, 2016.

Reuben A. Farrugia (S04, M09) received the first degree in Electrical Engineering from the University of Malta, Malta, in 2004, and the Ph.D. degree from the University of Malta, Malta, in 2009. In January 2008 he was appointed Assistant Lecturer with the same department and is now a Senior Lecturer. He has been in technical and organizational committees of several national and international conferences. In particular, he served. as General-Chair on the IEEE Int. Workshop on Biometrics and Forensics (IWBF) and as Technical Programme Co-Chair on the IEEE Visual Communications and Image Processing (VCIP) in 2014. He has been contriuting as a reviewer of several journals and counferences, including IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video and Technology and IEEE Transactions on Multimedia. On September 2013 he was appointed as National Contact Point of the European Association of Biometrics (EAB).

Christine Guillemot Christine Guillemot, IEEE fellow, is Director of Research at INRIA, head of a research team dealing with image and video modeling, processing, coding and communication. She holds a Ph.D. degree from ENST (Ecole Nationale Superieure des Telecommunications) Paris, and an Habilitation for Research Direction from the University of Rennes. From 1985 to Oct. 1997, she has been with FRANCE TELECOM, where she has been involved in various projects in the area of image and video coding for TV, HDTV and multimedia. From Jan. 1990 to mid 1991, she has worked at Bellcore, NJ, USA, as a visiting scientist. She has (co)-authored 15 patents, 8 book chapters, 50 journal papers and 140 conference papers. She has served as associated editor (AE) for the IEEE Trans. on Image processing (2000-2003), for IEEE Trans. on Circuits and Systems for Video Technology (2004-2006) and for IEEE Trans. On Signal Processing (2007-2009). She is currently AE for the Eurasip journal on image communication and member of the editorial board for the IEEE Journal on selected topics in signal processing (2013-2015). She is a member of the IEEE IVMSP technical committees.

REPLY TO REVIEWERS

The authors would like to thank the reviewers for their work and recommendations. We believe that their questions and recommendations helped improve the quality of the revised paper. Below we address all the queries raised and modified the paper accordingly. In summary, we have completely rewrote section IV, compared our method with Yang *et. al*. [76], derived results on an additional dataset (CAS-PEAL - as indicated by the previous reviewing board and mentioned by Reviewer 2), compared our method in the wild with the extended LINE and Eigen-Patches in Fig. 11 and included all references recommended by the reviewers. We would like to emphasize here that this work provides the most extensive evaluation of face hallucination where apart from evaluating it in terms of PSNR, we are also presenting recognition performance objectively using different face recognition systems. We also show that this method can perform very well in the wild - outperforming both generic SR methods and the extended LINE and Eigen-Patches (extended using the method in V).

Reviewer 1

1. This paper presents a two-step framework for face hallucination. The first step is to obtain a global optimal approximation which may lack some texture details. Then in the second step, the missed details are compensated through sparse support based ridge regression. Through experiments illustrate that: (1) The objective image quality scores is not always consistent with the perceptual quality of reconstructed face images and also the face recognition accuracy. Actually, this point has been also found on other literatures, such as Wang et al. "Evaluations on Synthesized Face Sketches", Neurcomputing, 2016. Related literature should be cited. (2) The proposed method could achieve state-of-the-art performance on the face images in the wild.

The authors would like to thank the reviewer for suggesting this very recent reference which we have included in the revised version (see sections I) and which further substantiates our conclusions. Note that, the recommended paper has been published in Nov. 2016 while ours was submitted in June 2016. This recommended paper was cited in reference [59].

2. In addition, the proposed two-step framework is very similar with another work on heterogeneous image transformation (Wang et al. Heterogeneous Image Transformation, PRL, 2013.) and the regression framework in this work is very related to the cascaded linear regression based image super-resolution work (Hu et al. SERF: A Simple, Effective, Robust and Fast Image Super-Resolver from Cascaded Linear Regression, IEEE TIP, 2016).

Once again, the authors would like to thank the reviewer for pointing two papers which have some level of similarity to our proposed scheme, and therefore were included in the introductory part of the paper to make the literature covered in this paper more complete (see [26] and [10]). Nevertheless, while some similarity exist, there are substantial differences between our proposed method and these papers listed here:

For example, our method differs from the "Heterogeneous Image Transformation":

1) The support in [26] is found on the low-resolution manifold which as found in [28], [40]–[42] is distorted because of the one-to-many mappings between the low- and high-resolution patches. This part is actually similar to the work of [34] who does exactly that for face hallucination (e.q. 3). In our case we are first finding a first estimate of the up-scaling function using all elements in the dictionary to derive the first approximation of the up-scaling function using (7) and use it to find a point on the high resolution manifold using (8). Then we exploit the locality of the first approximated solution to fund the support on the high-resolution manifold *i.e*. the atoms which are optimal to reconstruct the first approximation $\widetilde{\mathbf{y}}_i^{s\{0\}}$. Instead of using the weights directly, we use the k atoms whose weights have the largest value as support and we extract the coupled sub-dictionaries $\mathbf{L}_i\left(\mathbf{s}_i\right)$ and $\mathbf{H}_i\left(\mathbf{s}_i\right)$ to derive a refined up-scaling function using (14) where the final solution is computed using (15).
2) Unlike our approach, this method reconstructs the transformed (in our case super-resolved) patch i_j^2 using the notation in this paper using a weighted combination of atoms found using sparse coding on the initial manifold (in our case low-resolution manifold). In our case, we are modelling the up-scaling function by first finding and estimate and then refine it by exploiting the high-resolution manifold, as mentioned above.
3) The main similarity between this approach and our method resides in the use of regression where here they use support vector regression. Nevertheless, this approach considers the second layer as a de-blurring operator to restore missing detail and does not use the original input for up-scaling (as done in our case). In the referenced work they train an SVR model using the output of the first stage (which uses a weighted combination of atoms found from the low-resolution manifold) and uses the high-resolution examples to learn the regression model. In our case we are using multivariate ridge regression to model the up-scaling functions in (7) and (14).

Our approach differs from "A Simple Effective, Robust and Fast Image Resolver from Cascaded Linear Regression":

1) This work adopts a cascade of linear regression models to compute generic super-resolution. This method learns a sequence of regressors to degrease the mean square error with respect to the high resolution patches at each layer. While

the learning of the up-scaling function of the proposed method is similar to the first layer of the sequence of regressors used in this paper, the aim of the second layer in our approach is to find the support (and thus coupled atoms) on the high resolution manifold, and the next stage we only use atoms with this coupled support to learn the second mapping relation (see (14)) which better preserves the texture detail.

2) The paper mentioned here divides the dictionary using clusters computed using k-means on the t-th level dictionary. Therefore, the atoms selection process considered here is different from the one adopted in our approach where sparse coding is used to find the support. In the new version we show that sparse coding derives atoms that are more suitable to reconstruct $\widetilde{\mathbf{y}}_i^{s\{0\}}$ than k Nearest Neighbours. Moreover, instead of trying to reduce the error, we use the same dictionary where only the selected coupled support are used in the second layer - therefore our approach is not a cascade of regressors.
3) Please note that, the recommended paper has been published in Sep. 2016 while ours was submitted in June 2016.

While two of the papers mentioned were published after our submission in June, we appreciate the fact that the reviewers have included them since it ensures that the most relevant and recent papers are included in this new version of the manuscripts.

Reviewer 2

The authors propose an improved method for face image super-resolution. The proposed method uses an LM-CSS algorithm to learn the optimal up-scaling function and then exploit the geometrical structure of the high-resolution images from the high-resolution dictionary. However, my main concern for this paper is lack of novelty and I have some questions and unclear things in the proposed energy model.

1. Although the first step of learning the up-scaling function is somewhat new, the second layer that exploiting the locality of the high-resolution images to refine the first estimation using the high-resolution dictionary is based on some existing methods.

To address this remark we have restructured Section IV where we try to explain at high level the objective of each layer. We then placed the detail of each layer in two separate sub-sections.

We first of all thank the reviewer for stating that learning an up-scaling function as a first step is somewhat new. In order to emphasize the contribution provided by *Layer 0*, we have obtained additional results which clearly show that searching for the closest neighbour on the high-resolution manifold using the first approximation provided by *Layer 0* gives a better neighbourhood preservation and also reduces the reconstruction error , than when searching for the neighbourhood on the low-resolution manifold. We refer the reviewer to both Fig. 4 and Fig. 5 in the revised version of this paper. All face hallucination schemes (except the LINE method) try to find neighbours or sparse support on the low-resolution manifold. These results clearly show that computing the first approximation $\widetilde{\mathbf{y}}_i^{s\{0\}}$ using the first layer is beneficial, and allows to derive neighbours which are more coherent with the actual neighbours of the ground-truth we are trying to approximate. The LINE method first finds nearest neighbours on the low-resolution manifold using $k = 150$ neighbours and tries to improve the neighbourhood selected on the high-resolution manifold. However, since the first approximation is based on neighbourhood computed on the low-resolution manifold n_L, the neighbour preservation is inferior to our *Layer 0* (np_H) approximation. We therefore claim that LINE is not guaranteed to converge - and actually results in Fig. 11 and quality analysis in Table IV confirm this comment.

However, we believe that the second layer was not properly explained in the previous version of the paper - and this point was also remarked by reviewer 3. This second step is complementary, and goes together with the first step. One key issue provided by the second layer is to identify the support $\mathbf{s}_i$ (marked in orange in Fig. 7) which are optimal to reconstruct the first estimated solution $\widetilde{\mathbf{y}}_i^{s\{0\}}$. The size of the support is controlled by the parameter k which as shown in Section III has an impact on both PSNR and texture consistency for Neighbour Embedding. Therefore, the objective here is to find a support $\mathbf{s}_i$ which is optimal to reconstruct $\widetilde{\mathbf{y}}_i^{s\{0\}}$ that is formulated in equation (12). Instead of using the weighted combination using the derived $\boldsymbol{\eta}_i$, which will converge to the first approximation and thus will lack texture details, we use the support $\mathbf{s}_i$ to extract the k column vectors from $\mathbf{L}_i$ and $\mathbf{H}_i$ to get the sub-dictionaries $\mathbf{L}_i(\mathbf{s}_i)$ and $\mathbf{H}_i(\mathbf{s}_i)$. Here $\mathbf{s}_i$ corresponds to the k atoms with the largest magnitudes $\boldsymbol{\eta}_i$. These sub-dictionaries are then used to derive a refined up-scaling function $\boldsymbol{\Phi}_i^{\{1\}}$ using(14).

Given that only k atoms are used to derive the refined up-scaling function $\boldsymbol{\Phi}_i^{\{1\}}$ in (14), the solution is expected to get better texture consistency (k is much smaller than the total number of atoms available) while achieving higher PSNR values (the sparse support is derived using BPDN to get the sparsest solution which minimizes the least squares distance from the first approximation (12). Therefore, the resulting refined up-scaling function $\boldsymbol{\Phi}_i^{\{1\}}$ should provide a compromise between texture consistency and PSNR.

We believe that the second layer is innovative because

1) we seek for the sparse support on the high resolution manifold instead of on the low-resolution manifold or nearest neighbours on the high-resolution manifold (as done by LINE). The results in section IV-B show that the sparse support is a better reconstruction model and is able to find the k support that are optimal to reconstruct the first solution. Then

we exploit the locality of $\mathbf{\Phi}_i^{\{1\}}$ which is close to the ground truth in Euclidean space to derive a refined up-scaling function $\mathbf{\Phi}_i^{\{1\}}$ using (14).

2) the method is not iterative and is guaranteed to converge to an optimal solution. The experimental results demonstrates that it is much more stable than LINE when tested on two frontal datasets using 1830 images and on the non-frontal images from the IJB-A dataset.

So, finding the support on the high-resolution manifold and using them to refine the projection matrix is innovative and gives an edge over the LINE and other methods considered in the experiments in terms of both recognition and quality (objective and subjective).

2. Some symbols are not clearly defined in this paper. For example, x_i^s in (2), what does the superscript s mean?

The symbol x_i^s was defined prior equation (1) and is defined as a standardized low-resolution patch. Nevertheless, the authors agree that it is more appropriate to define this symbol following equation (2) where it is first used. This symbol was re-defined after equation (2) where we specifically mention that upper-script s stands for standardized vectors (zero mean unit variance).

3. The authors claim that PSNR does not correspond to the texture details based on some experimental results. Their response to Reviewer1 is also based on observation in results, but it is still not clear why PSNR always increase while texture details keep disappearing in Figure 2. Is there any theoretical explanation for this claim?

We have used a recommendation provided by reviewer 3 to show the LBP spectrum to show what is happening. Given that the original Fig. 2 was not giving enough information to the reviewers we decided to modify it such that apart from the facial image at different neighbourhood size, we also give the LBP spectrum and also PSRN and TC quality metrics. From the results in Fig. 1 it can be seen that the PSNR saturates at around $k = 200$ while texture consistency starts going down at that point for Neighbour Embedding. The reason why PSNR increases (although marginally) beyond this point is because it Neighbour Embedding employs a least squares solution which tries to minimize the MSE - and the MSE is smaller when increasing the number of column-vectors that are combined. Please note that unlike generic super-resolution, the textures combined are similar *i.e*. right eye patches are combined to restore the right eye patch of the low-resolution test image. However, if we add the number of column vectors (atoms) to combine, these will derive blurred patches, because patches with different textures (different location of the eye center, different eye brows, different expression) are being combined. If we select the k atoms which are closest in some sense *i.e*. all patches have their eyes open, then the reconstructed patch will have higher texture consistency.

In Fig. 2 we show the LBP spectrum and show that the LBP spectrum of faces reconstructed using $k = 200$ is closer to the LBP spectrum of the high-resolution image, which explains the higher TC metric. On the other hand, images reconstructed using $k = 1203$ are blurred and thus have more repetitive texture - which provides spikes in the spectrum. Moreover, given that these images contain less texture, the resulting spectrum is more sparse. Given that since the texture consistency metric measures the similarity between the low- and high-resolution spectrum, and that one can see that the spectrum of those images super-resolved using $k = 200$ are more consistent to that of the ground truth to those reconstructed using $k = 1203$, one can expect that the texture consistency metric starts decreasing if the number of atoms used to estimate a patch is too large.

The theoretical justification which explains why the blurred images using $k = 1203$ is achieving higher PSNR values is that it is based on the least squares and is therefore biased towards over smoothed (*i.e*. blurred) images. Several research efforts (see work of A.C. Bovik and many others) have tried to design metrics which do not favour blurred images. The texture consistency metric measures how the texture is related to the original texture of the original image. Given that the LBP spectrum is more noisy when using larger neighbourhood sizes, this will contribute to reducing the texture consistency, as seen in Fig. 1. Please note that these results were obtained using the neighbour embedding scheme [61], which finds the weighted combination on the low-resolution manifold and then uses the same neighbourhood and weights on the high resolution dictionary to restore the original image. Therefore, what is happening there is that increasing the number of neighbours is combining more facial images in the high-resolution images thus converging more to a blurred face image with lower texture. Therefore, the texture consistency will drop once more than 200 neighbours are considered.

4. The authors also do not provide compelling results that convince me it works substantially better than previous methods. As shown in Tables 1 and 2, the proposed method only achieves PSNR gains about 0.1-0.3dB.

Based on the recommendation of the same reviewer, we have tested our algorithm and compared to several other schemes on a different dataset - the CAS-PEAL dataset which was used by Jiang et al [47]. We have also added the face hallucination of Yang *et. al*. for comparison. It must be mentioned that unlike other face hallucination schemes where they only test on a small set of images (40 images in [47]), we choose two datasets where the gallery images are captured in controlled

environment while the probes are captured either outdoors (FRGC) or using different expressions for the CAS-PEAL dataset. We have evaluated our scheme using a total of 1830 probe images for recognition and quality evaluation. We must also mention that all face hallucination schemes just give results about PSNR on much smaller datasets and give no indication of its impact on recognition. We can therefore state here that the evaluation provided in this paper is the most extensive evaluation of face hallucination found in literature.

Our proposed LM-CSS achieves rank-1 recognition rates between 20% and 60% over bi-cubic interpolation. We are also outperforming a very-low-face-recognition scheme DFD published in PAMI in 2014 at very low-resolution, achieving gains between (34% and 60%) - thus showing for the first time that hallucination followed by recognition can be more advantageous at very low resolutions such as $d_x = 8$ where DFD achieves are rank-1 recognition of 2% while our LM-CSS followed by LBP face recognizer achieves rank-1 recognition of 56%. Apart from that we are outperforming Eigen-Patches (which is found to provide the most competitive PSNR) by achieving rank-1 recognition gains between 2% and 8%. The most competitive method in terms of recognition was found to be LINE followed by LBP face recognition, where we consistently outperform it by around 2% at different magnification factors on the CAS PEAL dataset. All this information is included in the discussion in the new version of this paper.

Apart from analysing the recognition performance, we did a quality analysis and found out that on the CAS-PEAL database we can achieve PSNR gains of around 1dB over LINE - which is the most competitive scheme in terms of recognition. On the other hand we are achieving like 0.3dB PSNR gain over Eigen-Patches on the FRGC dataset while the results on the CAS PEAL dataset are closer. Nevertheless, the subjective results in Fig. 11 clearly show that the images reconstructed using LINE are generally very noisy while images reconstructed using Eigen-Patches are generally very smooth. On the other hand, those images reconstructed using LM-CSS have more texture detail and of significantly better quality. It must be mentioned that if we want to achieve the highest PSNR values we can set $k = 1203$ where $\widetilde{\mathbf{y}}_i^{s\{1\}} = \widetilde{\mathbf{y}}_i^{s\{0\}}$ (see results in tables VI and VII in the reply to reviewer 3 Q.1, which confirms what I am saying here. However, this was not included in the paper because achieving the highest PSNR is not the aim of this work - but the aim is to get the facial images with the best subjective quality (as shown in Fig. 11) and can achieve the highest recognition rates.

5. Some raised questions by the reviewers are not answered by the authors. The reviewers asked the authors to compare the proposed algorithm with other methods on a new dataset, and compare with the work: Yang et al. But the authors did not consider these suggestions.

The proposed algorithm was compared to the method of Yang et al. and all algorithms were evaluated in terms of recognition and PSNR using the CAS-PEAL data base - which was recommended by the reviewer this reviewer is referring to. Our proposed method outperforms the method of Yang et al quite significantly. Moreover, the results on the CAS-PEAL dataset are quite consistent with those computed on the FRGC dataset. In fact, we can say that the including the CAS-PEAL dataset consolidates the superiority of our proposed scheme in terms of both quality and recognition.

Reviewer 3

This paper presents a two-step face hallucination method. The first step is to approximate HR face images by a linear model and then the second step is to enhance the facial details by a coupled LR/HR dictionaries. Since using the weights calculated from LR inputs to hallucinate HR faces needs to combine a number of HR faces, which may introduce distractions, this paper estimates an interpolation function to upsample LR inputs by a linear model. The paper assumes that in this case the upsampled HR faces are close enough to the ground-truth HR faces on the HR face manifold but they are still blurred. Hence, the paper tries to find k nearest neighbors to generate the details. Hence, they can achieve high-frequency details.

To be more accurate we try to find the k-sparse support approximation on the high-resolution manifold rather than the closest neighbours. We have re-structured Section IV to make the explanation of the proposed method more clear. Moreover, we include more results which show that the first step is essential to find an optimal approximation which preserves the local neighbourhood better. Moreover, the second layer is used to find the k-sparse support suitable to approximate the first approximation and use them to refine the up-scaling function that is able to recover texture detail which is more consistent to that of the ground truth.

This paper is well written and the methodology is clearly stated, which is easy to understand. However, in my opinion, the novelty is minor and in the experimental part the comparison is not sufficient/unfair, such as Fig. 8. Here are my concerns:

The authors would like to thank the reviewer for remarking that the paper is well written and that the methodology is clear. As mentioned, we have explained in more detail the function of each layer where the contribution and novelty component of each layer is better emphasized. Moreover, we have tested our proposed method against state of the art methods found in literature (7 hallucination methods and 1 very low resolution face recognition method) on two data-sets since we added

the results on the CAS-PEAL dataset. We hope that we will be able to adequately answer to the concerns raised by this reviewer.

1. I agree that higher PSNR does not mean better visual quality or better recognition accuracy. But in Part VI-B, why do the authors choose different k values? In my opinion, this paper should find a trade-off k which can achieve better PSNR while improving recognition accuracy. Furthermore, I am wondering what the differences would be given different values of k? I think the comparison is not fair because for different tasks the paper tries to use different parameters. From Fig.1 b, TS begins to decrease after 200, why do the authors use k=50 for recognition tasks? Why not set k=150 for all the experiments?

Please note that the results in Section III are derived using Neighbour Embedding scheme in [61], and is used to get a better understanding on the effect of k on the performance of neighbour embedding - which is the basis of most recent work on face hallucination. From these results we show that improving PSNR does not mean you get images which are coherent with the ground-truth in terms of texture which is more related to recognition. We introduce LM-CSS in section IV. Therefore, the results which show that TC decreases after $k = 200$ is for neighbour embedding and not for LM-CSS. In the revised version we make this more clear in section III and mention it again in section VI-A, where we are trying to find the optimal parameters for LM-CSS. The results in section VI-A show that the optimal parameters for LM-CSS is dependent on the application scenario. From these results it can be seen that the best performance in terms of recognition rate was achieved using $k = 50$ while adequate PSNR performance can be achieved using $k = 150$ and we decided to use this configuration for the results on both FRGC and CAS-PEAL datasets. Therefore, given these results it is not possible to find one optimal value of k for both scenarios. In fact, if you look at the PSNR results in tables III and IV Eigen-Patches, which uses all elements in the dictionary attains the most competitive performance to our scheme. However, even thou it achieves good PSNR performance it attains very low recognition performance (see results in table I and II). On the other hand, the LINE method, which employs k atoms from the dictionary, seems to be more competitive in terms of recognition but suffers in terms of PSNR. While we still outperform LINE in terms of recognition, our method achieves PSNR gains of up to 1dB using the same neighbourhood size *i.e.* $k = 150$. Therefore, it is not possible to derive a k optimal for both application domains.

To get a compromise with the point raised by the reviewer we include results with LM-CSS for both $k = 50$ and $k = 150$ in tables I - IV in the new version of this paper. Some additional results are included below to help the reviewer understand the effect that k has on performance. As one can see in Tables VI and VII, we show quality results in terms of PSNR and SSIM using different values of k for both Eigen-Patches, LINE and our proposed LM-CSS. It can be seen that even if we reduce the value of k to 50 for LM-CSS. the results are still superior to the performance of LINE when using $k = 150$ (optimal setting proposed in [47]). It must be mentioned that reducing the value of k for LINE makes their method unstable and provides significant losses in PSNR (up to 2dB relative to LINE with $k = 150$). On the other hand, our LM-CSS method is guaranteed to converge even when using a small number of support, such as $k = 50$. We must also mention here that to achieve the best performance in terms of PSNR we can set $k = 1203$, *i.e.* using all elements in the dictionary. However, setting $k = 1203$ produces blurred images which will reduce the recognition performance. Given that achieving the highest PSNR is not the prime aim of this work we decided not to include results with $k = 1203$ in this paper.

TABLE VI: Summary of the Quality Analysis results using the PSNR and SSIM quality metrics on the FRGC dataset.

	Resolution d_x							
Hall^n Method	**8**		**10**		**15**		**20**	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
Eigen-Patches [22]	27.3918	0.7778	28.3847	0.8196	30.3118	0.8842	31.8986	0.9203
LINE [47] ($k = 50$)	25.6811	0.6869	26.6226	0.7355	28.6833	0.8264	30.3465	0.8808
LINE [47] ($k = 150$)	27.0927	0.7591	28.0253	0.8009	30.0471	0.8727	31.6970	0.9131
Proposed ($k = 50$)	27.1307	0.7679	28.1078	0.8093	30.0240	0.8761	31.6875	0.9139
Proposed ($k = 150$)	27.4866	0.7802	**28.4200**	0.8009	30.3431	0.8845	**31.9610**	0.9209
Proposed ($k = 1203$)	**27.5140**	**0.7947**	28.4044	**0.8300**	**30.3431**	**0.8915**	31.9393	**0.9245**

In tables VIII and IX we show the recognition results using different configurations of k for Eigen-Patches, LINE and our method when using LBP face recognition. It can be seen that increasing the value of k to 150 for LM-CSS provides performance which is still competitive with LINE ($k = 150$), where actually we achieve the best performance on the CAS-PEAL dataset with $k = 150$ while we have performance very close to LINE on the FRGC dataset. It is also interesting to notice that the LINE method is very unstable when you reduce the neighbourhood size k to 50, where it achieves

TABLE VII: Summary of the Quality Analysis results using the PSNR and SSIM quality metrics on the CAS-PEAL dataset.

Hall^n Method	**Resolution d_x**							
	8		**10**		**15**		**20**	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
Eigen-Patches [22]	26.0834	0.8071	27.6536	0.8620	30.6109	0.9303	33.1584	0.9615
LINE [47] ($k = 50$)	22.7275	0.6176	24.2920	0.7018	27.3511	0.8349	29.8894	0.9039
LINE [47] ($k = 150$)	25.2454	0.7567	26.7365	0.8176	29.8449	0.9085	32.4446	0.9492
Proposed ($k = 50$)	25.5597	0.7704	27.0966	0.8318	30.0366	0.9125	32.6052	0.9521
Proposed ($k = 150$)	26.0875	0.8046	27.6003	0.8579	30.6199	0.9287	33.2108	0.9613
Proposed ($k = 1203$)	**26.1637**	**0.8122**	**27.6838**	**0.8638**	**30.6712**	**0.9319**	**33.2487**	**0.9627**

recognition rates even lower than Eigen-patches. This can be explained since the first approximation uses k neighbours on the low-resolution manifold to get the first estimate, this will be inaccurate since the neighbour preservation will be quite low. Therefore, LINE will find it very hard to converge to a good solution.

TABLE VIII: Summary of the Rank-1 recognition results and Area Under Curve (AUC) metric using two different face recognition algorithms (on the FRGC dataset.

Hall^n Method	**Rec^n Method**	**Resolution d_x**							
		8		**10**		**15**		**20**	
		rank-1	**AUC**	**rank-1**	**AUC**	**rank-1**	**AUC**	**rank-1**	**AUC**
Eigen-Patches [22]	LBP	0.5226	0.9625	0.6215	0.9704	0.7237	0.9800	0.7602	0.9830
LINE [47]($k = 50$)	LBP	0.5786	0.9604	0.6224	0.9696	0.7041	0.9785	0.7553	0.9821
LINE [47]($k = 150$)	LBP	0.5925	0.9647	0.6559	0.9714	0.7323	**0.9804**	0.7677	**0.9833**
Proposed ($k = 50$)	LBP	**0.6032**	**0.9658**	**0.6581**	**0.9722**	**0.7398**	0.9798	**0.7742**	**0.9833**
Proposed ($k = 150$)	LBP	0.5452	0.9644	0.6344	0.9710	**0.7398**	0.9801	0.7602	0.9831

TABLE IX: Summary of the Rank-1 recognition results and Area Under Curve (AUC) metric using two different face recognition algorithms (on the CAS-PEAL dataset.

Hall^n Method	**Rec^n Method**	**Resolution d_x**							
		8		**10**		**15**		**20**	
		rank-1	**AUC**	**rank-1**	**AUC**	**rank-1**	**AUC**	**rank-1**	**AUC**
Eigen-Patches [22]	LBP	0.5400	0.9192	0.6567	0.9497	0.8222	0.9726	0.8611	0.9708
LINE [47]($k = 50$)	LBP	0.3878	0.8673	0.5189	0.8895	0.7556	0.9492	0.8333	0.9699
LINE [47]($k = 150$)	LBP	**0.5822**	0.9220	0.6833	0.9458	0.8167	0.9744	0.8633	0.9797
Proposed ($k = 50$)	LBP	0.5622	0.9282	**0.6922**	0.9538	0.8256	0.9756	0.8689	0.9806
Proposed ($k = 150$)	LBP	0.5344	**0.9385**	0.6789	**0.9611**	**0.8322**	**0.9783**	**0.8756**	**0.9820**

2. As mentioned in the second contribution of this paper, previous methods are not acceptable as criminal evidence, because the hallucinated faces are combined from multiple HR faces or patches. Hence, this paper proposes a linear model to interpolate LR faces first. However, in the second step, this paper also needs to use multiple reference HR patches to hallucinate HR details, which are cropped from multiple HR faces as well. So my question is that if the hallucinated details also come from multiple facial parts, why the results generated by the proposed method are acceptable for forensic evidence?

We believe that the contribution of the second layer was not well presented in the previous version. As mentioned earlier, we decided to restructure the section of the proposed method to make the contribution of each layer more clear and make sure that the novelty component of each layer is more apparent. We have also added more results to show what is the gain provided by each layer. In fact, in both steps we are modelling an up-scaling function using a set of face images to derive a linear up-scaling function for each patch. The second step is complementary, and goes together with the first step. One key issue provided by the second layer is to identify the support $\mathbf{s}_i$ (marked in orange in Fig. 7) which are optimal to reconstruct the first estimated solution $\widetilde{\mathbf{y}}_i^{s\{0\}}$. The size of the support is controlled by the parameter k which as shown in Section III has an impact on both PSNR and texture consistency for Neighbour Embedding. Therefore, the objective here is to find a support $\mathbf{s}_i$ which is optimal to reconstruct $\widetilde{\mathbf{y}}_i^{s\{0\}}$ that is formulated in equation (12). Instead of using the weighted combination using the derived $\boldsymbol{\eta}_i$, which will converge to the first approximation and thus will lack texture details, we use the support $\mathbf{s}_i$ to extract the k column vectors from $\mathbf{L}_i$ and $\mathbf{H}_i$ to get the sub-dictionaries $\mathbf{L}_i(\mathbf{s}_i)$ and $\mathbf{H}_i(\mathbf{s}_i)$. Here $\mathbf{s}_i$ corresponds to the k atoms with the largest magnitudes $\boldsymbol{\eta}_i$. These sub-dictionaries are then used to derive a refined up-scaling function $\boldsymbol{\Phi}_i^{\{1\}}$ using (14). Therefore, $\boldsymbol{\Phi}_i^{\{1\}}$ is still an up-scaling function which does not combine patches from different faces, and is expected to be more accepted by the forensics community.

3. As for the third contribution, LM-CSS uses manually labeled facial landmarks to align reference HR faces. In my opinion, there is no big difference from existing methods [12-36, 38-41]. Because the manually labeled landmarks can be easily embedded to the state-of-the-art methods, I think the novelty of the third contribution is really limited.
5. In Fig. 8, this paper mainly compares with general super-resolution methods rather than the state-of-the-art face hallucination methods, such as [30, 31, 44]. I think Fig.8 should demonstrate the visual results of those methods.

The authors agree with the fact that the third contribution can be applied to a number of face hallucination methods including LINE, Eigen-patches etc. Nevertheless, the authors of LINE claimed they are doing face hallucination on real-world images and poor performance was obtained when the face image deviated slightly from frontal (see last row of Fig. 13 of reference [47]). Given that the method proposed for the third contribution is extendible to other schemes does not make it limited but actually makes it more widely usable. It shows that by aligning the dictionaries to the input low-resolution test image, one can employ face hallucination in the wild. Such scheme was never published which makes it novel while not being complex. On the other hand, schemes based on deep learning which are said to be able to compute face super resolution in the wildn(will go into more detail when answering question 6) are not that good since they have a number of limitations mentioned below. We here demonstrate the images restored using LINE, Eigen-Patches and LM-CSS with non-aligned dictionaries (see Fig. 13. Of course the quality of the reconstructed faces is poor because they are not aligned - but that is their actual implementation. On the other hand, better quality can be achieved using generic super-resolution schemes such as NCSR and SRCNN (deep learning based super-resolution scheme).

To provide a more adequate answer to question 5 we have integrated Eigen-patches and LINE in our dictionary alignment strategy and show the results in the newer version (see Fig. 11). It can be immediately noticed that wile the hallucination methods provide sharper face images, the images reconstructed using Eigenpatches with aligned dictionary are blurred while those reconstructed using the aligned LINE method have structural distortion - which is consistent with the results presented on frontal images. On the other hand the images constructed using aligned LM-CSS are of much better quality at a magnification factor of 5 starting from a resolution where inter-eye distance is of only 8 pixels.

4. This paper employs LBP features to measure the texture similarity. I may suggest that by showing image spectrum of hallucinated faces readers could easily tell what kind of frequency components are hallucinated rather than a value of TS. From the spectrum, it also can be easy to tell whether the generated high-frequency details are consistent with the ground-truths or not.

We would like to take the opportunity to thank the reviewer for this suggestion. In this revised version we modify Fig. 2 to use the suggested spectrum of LBP to show what happens when we increase the neighbourhood size k when using neighbour embedding. Using the LBP spectrum one can notice that while the face image is more blurred ($k = 1203$) the resulting spectrum has more noisy spikes which contributed to reducing the texture consistency when using larger neighbourhood size. We believe that the use of this spectrum in Fig. 2 helps the reader to better understand what the texture consistency metric is doing and to explain why it drops when using larger neighbourhood sizes. Please note that the results in Fig. 2 are for neighbour embedding [61] and not for the LM-CSS.

6. Very recently, there are some deep learning based face hallucination methods have been proposed, the authors should not miss those.

The authors thank the reviewer to point us towards recent advances in deep learning. The method [50] adopts a bi-channel convlutional neural network to reconstruct Gaussian and Motion blur. They use a 48×48 input network, which makes it quite rigid for super-resolution. If for example, you have an image of resolution 20×20 one has to up-scale the

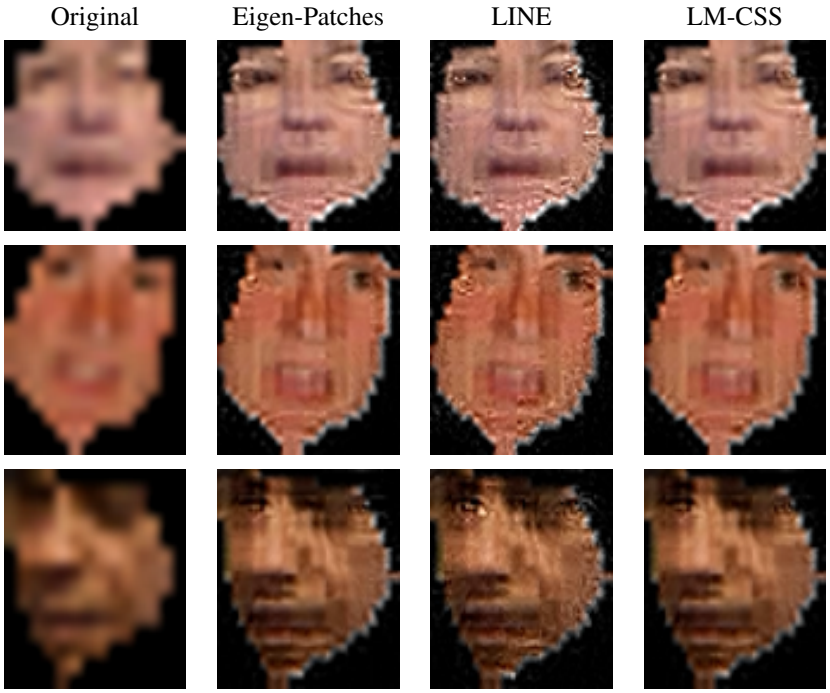


Fig. 13: Face hallucination in the wild without dictionary alignment.

input image to 48×48, while if you have a higher resolution you must down-scale the input image to 48×48, thus loosing information. In their results in Fig. 4 they consider images of resolution 50×50, which roughly corresponds to an inter-eye distance of $d_x = 30$. Please note that we are considering much lower resolutions in our work. Another important thing to mention, is that while they are claiming face super-resolution in the wild, the face images using in the evaluation are quite frontal. The results we present in Fig. 11 and 12 are much more realistic and truly in the wild. We have contacted to authors of this paper to give us the code or learned models - but they did not reply. Nevertheless, this contribution will be included in the introductory part of our new version of the paper (ref [**?**]).

Another project that one can find on github is the srez project https://github.com/david-gpu/srez. We have tried to implement their method - however when we contacted the developer of this project to see whether we can use their code to super-resolve new data, he answered as follows:

"Right, there current code does not support the user to provide a particular 16x16 image to be supersampled. *It's just a proof of concept prototype*. You would need to modify the source code to read an external image and upsample it."

The srez algorithm is very restrictive *i.e.* it can only super-resolve images of resolution 16×16 and achieves a magnification factor of 4. So it has to be retrained for different input resolutions and different magnification factors. Please note that in our work to perform face super-resolution in the wild we are not restricted by the input resolution or the magnification factor we want to achieve. Nevertheless, experimental results show that it it hard to reliably super-resolve facial images with an inter-eye distance lower than 8 pixels. Method based on deep learning is restricted by the number of input and output neurons.

We have done some experiments to see how good the srez project is: In fact if the reviewer has a look at the images in https://github.com/david-gpu/srez, one can immediately notice that while the reconstructed faces are appealing, they do not resemble the original face image. In fact, we found out that the PSNR of the facial region averaged over all images of the facial region is just 20.75dB - which is very low. This can be explained by the fact that the inter-eye distance of the low-quality image is just 5 pixels, which is too small and thus the low-quality image has too little information to exploit to reconstruct a reliable approximation of the high-resolution face image.

Base on the above concerns, I think the novelty of this paper is limited which cannot reach the bar of TIP right now. Hence my decision is major revision.

We hope the reviewer is more satisfied with the modifications and replies, and finds the contribution of this paper more appropriate for publication in TIP.