ASIE: Application Specific Image Enhancement for Face Recognition

Computational Biomedicine Lab, Department of Computer Science
University of Houston, Houston, TX USA

ABSTRACT

In this paper, we propose a novel method to enhance low quality images. Specifically, we focus on facial images. Low quality images are often degraded by motion artifacts, sensor limitations, and noise contamination leading to loss of higher order information that is essential for face recognition. First, we demonstrate that conventional denoising and deblurring methods are not able to fully recover the latent image resulting in residual artifacts in the image. Then, we present a novel approach for image enhancement that removes these residual artifacts using sparse encoding methods. The potential of the method is demonstrated through promising results on facial images for face recognition application.

Keywords: Face recognition, image enhancement, deblurring, denoising

1. INTRODUCTION

Advanced face recognition systems largely depend on higher order information present in the images for accurate recognition. The extraction of such higher order information depends on the spatial resolution and not just on the pixel resolution. The spatial resolution is defined as the number of independent pixels per unit length, whereas the pixel resolution is defined as the number of pixels per unit length. With the advancement of camera technology, the pixel resolution of acquired images has increased significantly. However, an acquired image may still exhibit lower spatial resolution due to the system noise or blurring caused by the sensor’s limitations and the object’s motion. For example, the images from surveillance cameras can be affected by both camera and subject motion resulting in surveillance images that are blurry and noisy.

Most of the existing face recognition algorithms are not designed to perform recognition when the input images are of low spatial resolution. This can be attributed to the fact that a linear blur model cannot account for the non-linear outliers that often exist in real imaging systems. These outliers generally fall in the categories of saturated pixels, non-Gaussian noise, nonlinear response curve of the cameras, and the residual artifacts that remain after denoising the image.

The commonly used approach of performing denoising and deblurring in succession provides only a partial solution to the problem of enhancement of the image quality. Image denoising in the presence of blur cannot fully resolve between the higher order information of the noise and that of the original image. Image deblurring is an ill-posed problem in the presence of noise and may accentuate the effect of the residual noise in the image. Therefore, even after performing denoising and deblurring of the acquired image, the resulting image still contains residual artifacts which hinders face recognition systems. Thus, there is a critical need to develop image enhancement methods that account for the residual artifacts.

In this paper, we present a novel application-specific image enhancement (ASIE) method that takes into account the structure of the object depicted for a specific application (e.g., faces in face recognition application). We first demonstrate that performing denoising and deblurring provides only an intermediate solution that still contains the residual artifacts. We propose a novel residual artifact removal step to obtain an enhanced image using sparse encoding methods. We formulate it as an example-based super resolution problem where prior knowledge about the imaged object (i.e., the face) is used. Such prior knowledge is obtained offline from the training examples using sparse representation techniques. Our main contributions are: (i) we demonstrate that the resulting images from the state-of-the-art denoising and deblurring methods exhibit residual artifacts that reduce the accuracy of face recognition systems, (ii) we present an application-specific image enhancement method that uses prior knowledge of the object structure in the images (modeled using the sparse representation techniques), (iii) we introduce an artifact removal step to reduce the residual noise in the denoised and deblurred images. We compared the performance of our method to state-of-the-art methods for the purpose of face recognition.
recognition. The evaluation was performed on a difficult dataset acquired from surveillance images on which a commercial face recognition software failed to operate. The results of the face recognition systems using our ASIE method demonstrate the superiority of our method over other image enhancement methods.

2. PREVIOUS WORK

Image noise and blur pose two of the most prominent challenges in the field of image processing and are generally studied under the topics of image restoration, image reconstruction, or deblurring/deconvolution. We study these problems keeping in mind the end goal of face recognition, which requires higher spatial resolution images. In this paper, we term the overall problem of constructing an application-specific higher spatial resolution image from the observed information as an image enhancement problem.

Blind deconvolution is an ill-posed problem, for which many approaches have been proposed. Many earlier approaches have used multiple images to estimate the blur. Cho et al. used multiple images with space-variant blurs. Rav-Acha and Peleg used images with different blurring directions to help reconstruct detail. Yuan et al. used combined information from blurred and noisy images. However, in practice, multiple images obtained within a constrained time interval may not be available.

Many other approaches operate on a single image to estimate the unobserved latent image. These approaches have to make assumptions on the form of the blur kernels because blind deconvolution with a single image is much more difficult than using multiple images. Dai and Wu performed alpha matting to estimate local blur kernels. Ji and Liu proposed a method based on spectral analysis of image gradients to identify blur kernels for general motion types. However, these approaches are not suitable to operate on images with more diverse 2D blur.

Fergus et al. proposed a variational approach using natural image statistics. To determine such complex statistical model is computationally expensive. Jia used transparency information from an alpha matte that is user-specified, and therefore it depends on the quality of the input alpha matte. Shan et al. formulated a MAP problem which is solved using an advanced iterative optimization technique. Their method is computationally expensive as it takes several minutes to deblur an image.

Recently, Cho et al. proposed a deconvolution method that can handle outliers. Their method requires non-blind deconvolution. Earlier, Cho and Lee proposed a fast blind deconvolution method for motion deblurring. Their method is very sensitive to initial parameter values and requires a couple of trial runs for each image to obtain the desired results. Krishnan et al. proposed a very effective blind deconvolution method with a novel regularization term that favors sharp images. However, it cannot eliminate the residual artifacts from lower resolution images.

3. METHODOLOGY

3.1 Notations

We refer to an image of low spatial resolution \( y \in \mathbb{R}^{m \times n} \) as low-resolution, and an image of high spatial resolution \( x \in \mathbb{R}^{m' \times n'} \) as a high-resolution image. The \( i^{th} \) patch of an image \( y \) will be denoted as \( y_i \in \mathbb{R}^{M \times N}, M \times N = 8x8x3, i = 1, \ldots, N_y \), where the total number of patches of \( y \) is denoted as \( N_y \). The dictionary for the noise removal is denoted as \( D_\eta \), and the dictionary for the artifact removal is denoted as \( D_\gamma \).

3.2 Formulation

In the usual formulation the acquired image \( y \) is the result of the high-resolution image \( x \) blurred by a matrix \( H \) along with the addition of a Gaussian i.i.d noise \( \eta \):

\[
y = Hx + \eta . \tag{1}
\]

In the above equation, the blur matrix \( H \) and noise \( \eta \) are the unknown variables, whereas the acquired image \( y \) is the only known variable. It is difficult to estimate \( x \), when there are many unknowns and only one known variable. We formulate this problem as an example-based super-resolution problem, where external information from the training images is used. Next, we describe step-by-step what is required to estimate the high-resolution image \( x \). If we subtract noise from both sides in Eq. 1, we obtain:

\[
y - \eta = Hx . \tag{2}
\]
Here, we assume that we know the variable for the noise, as well as the blurring operator. Multiplying the pseudo-inverse $H^\dagger$ of both sides of Eq. 2 we obtain:

$$H^\dagger(y - \eta) = H^\dagger H x .$$  

However, $H^\dagger H \neq I$, where $I$ is identity matrix. Let us denote $g$ as the artifact that remains after applying the pseudo-inverse:

$$H^\dagger H x = x + g .$$

By combining Eqs. 3 and 4 and subtracting $g$ from both sides we obtain:

$$x = H^\dagger(y - \eta) - g .$$

From Eq. 5 (Fig. 1), note that to estimate the high-resolution image $x$ we need to: (i) remove noise $\eta$, (ii) deblur the denoised image, and (iii) remove the artifacts $g$. If we treat $g$ as noise or Gibbs phenomenon,\(^{12}\) then the algorithm proposed by Yuan et al.\(^ {12}\) can be used to remove them. However, their algorithm smooths out the edges, which are important for the face recognition engines. We propose a different solution to remove these artifacts by using sparse encoding methods on the training images. For steps (i) and (ii) we used existing denoising and deblurring algorithms with modifications. In the following sections, we explain our method for each step.

### 3.3 Denoising

Many state-of-the-art algorithms have been proposed in order to reduce the noise contamination in images. In our problem, we are focusing on real natural images which do not exhibit strong noise contamination. Note that, if the noise is not removed in the first step before applying the pseudo-inverse of blur transformation (Eq. 5), then it will be amplified and will contaminate the neighboring regions to the extent that the noise will be assumed to be a part of the foreground image. Let $y^*_i$ denote a denoised image patch $y_i$:

$$y^*_i = y_i - \eta .$$

To solve this problem, we use a method proposed by Aharon et al.,\(^{13}\) but instead of building a dictionary from an input low-resolution image, we build it by using high-resolution images from the training database, which is computed offline. Let $y_i$ be represented as a linear combination of sparse vector $\alpha$ with respect to the dictionary $D^n$:

$$y_i \approx D^n \alpha_i ,$$

where $\alpha_i$ is a sparse representation of image patch $y_i$. In Eq. 7, given image patch $y_i$ and dictionary $D^n$, we want to represent $y$ as a combination of a few atoms of the dictionary with a weight parameter $\alpha_i$:

$$\min_{\alpha} \| \alpha \|_p \quad \text{s.t.} \quad \| y_i - D^n \alpha_i \|_2 < \xi ,$$

where $\| . \|_p$ is $l_p$-norm, and $\xi$ is a threshold value. Since we want to represent $y$ as a combination of a few atoms of $D^n$, then $p = 0$, where $l_0$-norm returns the total number of non-zero elements. This problem is a combinatorial problem (NP
It has been suggested that if $\alpha_i$ is sparse, then the condition $p = 0$ can be replaced with $p = 1$, which yields a Lasso problem.\cite{14,15} Setting $p = 1$ and using the Lagrange multiplier, Eq. 8 becomes:

$$
\min_{\alpha_i} \| y_i - D^p \alpha_i \|_2 + \lambda \| \alpha_i \|_1
$$

(9)

where $\lambda$ is a regularization parameter that controls the sparsity. The resulting image $y^*$ is estimated as $y^*_i = D^p \hat{\alpha}_i$, where $\hat{\alpha}_i$ is estimated using Eq. 9.

### 3.4 Deblurring

Let $h$ denote a blurring kernel of blur matrix $H$, then Eq. 1 can be written as:

$$
y = h \ast x + \eta ,
$$

(10)

where $\ast$ denotes the convolution operator. Our goal is to estimate the blur kernel $h$. Krishan et al.\cite{11} proposed to estimate the kernel by using high-frequencies of the image as follows:

$$
\hat{h} = \arg \min_{h, x_{1,2}} \| y_{1,2} - h \ast x_{1,2} \|_2 + \lambda \| h \|_1 + \frac{\| x_{1,2} \|_1}{\| x_{1,2} \|_2} ,
$$

(11)

where $y_{1,2}$ represents the high-frequency components of image $y$ in both the horizontal and vertical directions. The advantage of this algorithm over other existing blind deconvolution algorithms is the simplicity of the cost function formulation, which makes it robust and efficient, and guarantees convergence. Now, let $\tilde{y}$ represent the denoised and deblurred image. After estimating a blur kernel $h$, we can recover the resulting image using non-blind deconvolution as proposed in:\cite{16}

$$
\min_{\tilde{y}} \| y_1 \|_p + \| y_2 \|_p + \lambda \| h \ast \tilde{y} - y \|_2 ,
$$

(12)

where $y_1$ and $y_2$ are horizontal and vertical high-frequency components, and $p = 0.8$.\cite{16}

### 3.5 Artifact Removal

Once the image is deblurred after noise has been removed, Eq. 5 will be as follows:

$$
x = \tilde{y} - g ,
$$

(13)

where $\tilde{y}$ is the denoised and deblurred image, and $g$ denotes the artifacts that are the result of non-blind deconvolution. In recent years, many algorithms have been proposed to remove the artifacts. We propose to remove these artifacts using a sparse representation method similar to the one that is described in Sec 3.3. In Sec. 3.3, we used smaller patch sizes to remove the noise. In this step, we use a bigger patch size as the neighboring pixels are contaminated by leftover noisy pixels during the non-blind deconvolution (Fig. 2). Note that the acquired image as well as the training images are facial images as we are focusing on the face recognition problem. Given a patch $\tilde{y}_i$ of image $\tilde{y}$, where $i$ is the index of $i^{th}$ patch, the sparse representation $\beta_i$ with respect to the dictionary $D^g$ is estimated as follows:

$$
\min_{\beta_i} \| \tilde{y}_i - D^g \beta_i \|_2 + \lambda \| \beta_i \|_1 .
$$

(14)

Once the sparse representation $\beta_i$ is estimated, then the high-resolution image patch is recovered as $x_i = D^g \beta_i$. 
Algorithm 1 Dictionary Building

**Input:** \( x^j, j = \{1, \ldots, K\} \)

**Output:** \( D^g, D^\eta \)

1: Partition images randomly into \( Z \) groups
2: **for all** \( z = \{1, \ldots, Z\} \) **do**
3: **for all** \( x^k \in G_z \) **do**
4: \( x^k \in G_z \) \( \in R^{8\times8\times3} \), \( x^k \in R^{16\times16\times3}, i = 1, \ldots, N_\eta, j = 1, \ldots, N_g \)
5: **end for**
6: Build dictionaries \( D^g_z \) and \( D^\eta_z \) (Eqs. 8 and 14)
7: **end for**
8: Concatenate \( D^g = [D^g_1 \ldots D^g_Z] \) and \( D^\eta = [D^\eta_1 \ldots D^\eta_Z] \)
9: Set \( X^\eta \leftarrow D^\eta \) and \( X^g \leftarrow D^g_z \)
10: Build final dictionaries \( D^\eta \) and \( D^g \) from \( X^\eta \) and \( X^g \) (Eqs. 8 and 14).

4. IMPLEMENTATION DETAILS

In this section, we describe the implementation of the proposed framework in detail. First, we trained two dictionaries for denoising and artifact removal. A single training database of face images is used to build both dictionaries (Fig. 1). We extracted patches of size \( 8 \times 8 \times 3 \) for denoising \( (D^g) \) and of \( 16 \times 16 \times 3 \) for artifact removal \( (D^\eta) \) with overlap of 7 and 14 pixels, respectively. Learning the dictionaries using the overlapping patches presents a significant challenge due to its high computational memory requirements. For example, for a color image \( y \in R^{256\times256\times3} \), the total number of patches \( (y_i, i = 1, \ldots, N_\eta) \) extracted from this image would be \( N_\eta = 125 \times 125 \). Thus, for \( K = 2,000 \) images in the training database, the total number of patches would be \( K \times N_\eta \) requiring \( 8 \times K \times N_\eta \times M = 44 \) GB of RAM to build the dictionary for denoising in double precision. Similarly, it would need 75 GB of RAM to build the dictionary for artifact removal. Furthermore, at least twice the required memory size would be needed to perform computations on such large matrices.

Unlike previous approaches that randomly selected a fewer number of patches to reduce the computational time and memory, we used a two-stage patch selection method for dictionary learning. In the first stage, we partitioned the training database into \( Z \) groups \( (G_z, z = 1, \ldots, Z) \), and build the dictionaries for each group \( (D^g_z) \) separately. Then, we combined them together as \( D^g = [D^g_1 \ldots D^g_Z] \) and rebuilt the final dictionary \( D^g \). The same procedure is followed to construct dictionaries for both denoising and artifact removal, respectively. Many optimization algorithms have been proposed in the past to build the dictionaries.\(^ {17–21} \) We used the SPAMS toolbox\(^ {21} \) because of its simplicity and robustness. Algorithm 2 provides the steps to build the dictionaries from a large number of patches.

To reconstruct the input image \( y \) noise removal is performed first. To reduce noise, an image \( y \) is partitioned into overlapping patches \( y_i \in R^{8\times8\times3} \). Then, for each patch the sparse encoding method is applied as it is described in Sec. 3.3. In Eq. 9, the dictionary \( D^\eta \) that is built offline for denoising from training images is used. After each patch is reconstructed, the overlapping pixels are averaged. Blind deconvolution is applied to estimate the blur kernel, and then non-blind deconvolution is applied (Sec. 3.4). After the deblurring, the image \( \tilde{y} \) will have some artifacts. These are mostly observed in surveillance camera images. To remove these artifacts we partition image \( \tilde{y} \) into bigger patches than the patches used for denoising \( (\tilde{y}_i \in R^{16\times16\times3}) \), and for each patch its sparse representation \( \alpha_i \) is estimated with respect to the dictionary \( D^\eta \), where the final reconstructed patch is estimated as \( x_i = D^\eta \alpha_i \). As in the denoising step, the overlapping pixels are then averaged. Note that due to the averaging of the overlapping pixels the output image will be slightly smoother. If we extract non-overlapping patches then we will have discontinuity (block artifacts) between adjacent patches.

5. EXPERIMENTAL RESULTS

In this section, we describe the dataset that we used to test our method and other methods. Comparing the pixel intensity difference (MSE, PSNR) between images acquired from two different modalities can be misleading due to the groundtruth requirement for MSE and PSNR. Thus, to measure the performance of the methods we used face recognition performance as performance measure, since it computes the difference between the features of the images instead of pixel intensities. We compare our method against the methods proposed by Krishnan et al.\(^ {11} \) and Cho et al.\(^ {1} \)
Algorithm 2 Reconstruction

Input: \( y, D^\eta, D^g \)

Output: \( x \)

1. Extract patches \( y_i \in \mathbb{R}^{8 \times 8 \times 3}, \ i = 1, \ldots, N_\eta \)
2. for all \( i = \{1, \ldots, N_\eta\} \) do
   3. Estimate \( \alpha_i \) using (Eq. 9)
   4. Reconstruct the patch \( y_i^* = D^\eta \alpha_i \)
3. end for
4. Obtain \( y^* \) by averaging the values in overlapping regions of the patches \( y_i \)
5. Estimate the blur kernel (Eq. 11)
6. Deblur image \( \tilde{y} \leftarrow y^* \)
7. Extract patches \( \tilde{y}_i \in \mathbb{R}^{16 \times 16 \times 3}, \ i = 1, \ldots, N_g \)
8. for all \( i = \{1, \ldots, N_g\} \) do
   9. Estimate \( \beta_i \) using Eq. 14
   10. Reconstruct patch \( x_i = D^g \beta_i \)
11. end for
12. Obtain \( x \) by averaging the values in overlapping regions of the patches \( y_i \)

5.1 Facial Dataset

To evaluate the performance of the proposed method and other existing methods, we use challenging low-resolution images from a tracking surveillance camera. For this data (UHDB), LR video sequences were obtained from a tracking surveillance system that uses a wide-field camera in an uncontrolled-lighting indoor environment. The UHDB database consists of data from 32 subjects. The HR images were obtained using a DSLR camera in a controlled environment and were used as a gallery. Each subject was asked to attend two sessions and walk within the surveillance area. The tracking camera was tracking the subject and acquiring frames. We used PittPatt to detect facial regions from each frame. Facial images from the first session were used to train the dictionaries. From the second session, we selected those facial images where PittPatt could not extract signatures from facial image. PittPatt uses these signatures to perform recognition.

5.2 Face Recognition Systems

As the original high-resolution (HR) image for a given LR image is not available, we cannot use MSE or PSNR to evaluate the performance of the proposed method. Thus, we used several face recognition systems. We demonstrate that the proposed methods are not tuned for one particular face recognition system. We used two different FR systems, one of which is a commercial FR system. The first was FaceIt developed by L-1 Identity Solution, and the second was the Sparse-Representation-based FR introduced by Wright et al. Each of these FR systems computes the similarity between two facial images differently. Thus, the recognition rate can be different for these FR systems even for the same facial images.

5.3 Results

Our proposed algorithm contains novel approaches to both noise and artifact removal. We applied ASIE on selected images from the UHDB dataset to demonstrate its effectiveness. In total, we used 154 images for testing, and 2,840 images to train the dictionaries. Figure 3 depicts the recognition rate of the face recognition systems on images that were enhanced using our ASIE method and previous methods. Our method outperforms other methods by 11% and 2% in rank-1 recognition rates for FaceIt and Sparse Representation-based FR, respectively. In Fig. 4, we present the qualitative results of our ASIE method and other existing methods for few example images on which the face recognition system failed when other enhancement methods were used except for ours. Note that, the output image of the method proposed by Krishnan et al. is sharp but it still contains artifacts. The output image of the method proposed by Cho et al. contains less artifacts but it is severely blurred. On the other hand, our method has superior performance in removing the artifacts as well as the blur from the images.

Figure 5 depicts the qualitative results of the image enhancement methods for example subjects that were either correctly or incorrectly identified by the FR engine. When the image is acquired for a slow moving subject, each method...
Figure 3. Face recognition results after using the spatial resolution enhancement methods on the DB dataset. (a) FaceIt, and (b) Sparse Representation based FR. Note that when ASIE is used, the recognition rates are higher than two other existing methods.

Figure 4. Depiction of the image enhancement algorithms. (a,e) Input image, Output of (b,f) Krishnan et al.,\textsuperscript{11} (c,g) Cho et al.,\textsuperscript{1} and (e,h) ASIE.

is able to enhance the image for accurate identification by the FR engine. On the other hand, when the subject moves fast in the scene (more blurry) then none of the methods is able to enhance the acquired image to a sufficient level for identification by the FR engine. Note that our method provides qualitatively better reconstruction of faces compared to the other methods, even for the cases where FR system failed to recognize the subject. This can be attributed to the fact that our proposed artifact removal step not only removes the residual artifacts but also attempts to recover the latent image using the prior knowledge of facial structures encoded in the learned dictionary from training database of facial images.

6. CONCLUSION

In this paper, we demonstrated that the existing state-of-the-art image enhancement methods (denoising and deblurring) provide incomplete solution for enhancing images from surveillance cameras for the purpose of face recognition. We proposed a novel approach to enhance the quality of images acquired for a specific application. We take advantage of the prior knowledge about the object being imaged for a particular application (i.e., face recognition).

Towards this, we presented an application-specific image enhancement (ASIE) method with a residual artifact removal step that provides sufficiently enhanced resulting images to be use with existing face recognition systems. Our method is based on the intuition that prior knowledge of the object structure provides reliable information to eliminate the residual artifacts as well as to reconstruct the latent image. We modeled the prior knowledge using sparse-representation techniques on image patches. We demonstrated the superiority of our ASIE method over existing state-of-the-art methods through evaluation of the performance of two face recognition systems on low-resolution images obtained from real world surveillance cameras.
REFERENCES


