Enhancing the Resolution of Low-Quality Images with an Example-Based Learning Approach

Kevin Huynh
(704283105)

Xinyu Wang
(505029722)

Abstract

Even though technology is rapidly improving, many visual surveillance systems are equipped with sub-par cameras that can not be replaced due to a variety of reasons. Handling the low-resolution images that these equipment capture is of the utmost importance and is a common problem in many fields, such as medicine, astronomy, and meteorology. Because scientists can not analyze this data directly due to its poor resolution, a great deal of research and thought have been put into how to estimate or synthesize high-resolution images from either a single or series of low-resolution images. With these estimated high-resolution images, scientists will be able to extract much more information than possible with their original data.

In this project, we will completely reimplement Park et al.'s paper by using an example based, object-class-specific, and top-down method to build a model which takes low-resolution facial images as input and outputs enhanced high-resolution facial images [1]. Furthermore, we will extend their model to another similar dataset to test how applicable it is to other types of images.

1 Introduction

As the scientific field keeps making progress, there is a growing interest in improving surveillance system for various reasons, such as security. In reality, it is difficult to obtain high-resolution images or videos from these systems due to the variation of illumination combined with the abrupt, fast movement of objects. These difficulties often result in unrecognizable or low-resolution images. The problems are especially aggravated in face detection. Even with the help of advanced high-definition video recorders, face images may still suffer from distortion and occlusion.

Numerous studies have been conducted in order to improve the quality of low-resolution images. One of the most successful techniques is to estimate and synthesize high-resolution images from low resolution images. For our project, we aim to estimate the parameters which correlates to synthesis of high-resolution faces with a limited number of corresponding high-resolution and low-resolution face pairs.

The method which this project uses is an example based, object-class-specific and top-down approach proposed by Jeong-Seon Park and Seong-Whan Lee [1]. The method takes a combination of low-resolution and high-resolution faces to learn a morphable face model and then further improves the parameters of this model by using recursive back-error projection procedures.

The report is organized as follows. We will explain how the data is represented and the structure of the back-error projection procedures in the next two sections. Afterwards, the overall implementation of our resolution enhancement model will be described along with the technical details of its inner workings. Lastly, our results and any further findings will be discussed.

2 The Representation of a Face

Instead of using a raw face image as an input, we convert a face into its shape and texture information, a method proposed by Thomas Vetter and Nikolaus F. Troje [2, 3]. According to former research, when two faces are compared, the differences in surface properties are characterized as changes in texture, while the differences in the spatial arrangement of the object features are characterized as changes in shape. Therefore, any differences between two faces can be separated into two different components, namely, texture and shape.

In our project, the shape information of an image can be obtained by applying a gradient-based optical flow algorithm. The optical flow algorithm takes a face image and a reference face image as input, and returns a displacement vector field. Thus, if the input image is of size n * m, the output is a matrix S of size n * m * 2 for any 2D image. Each element in S(:,:,1) denotes the horizontal displacement of a pixel that corresponds to the same pixel of the reference image. Similarly, Each element in S(:,:,2) denotes the vertical displacement of a pixel that corresponds to the same pixel of the reference image. The shape information is thus a 3D matrix.

The texture information of an image is obtained by back warping. According to Jeong-Seon Park and Seong-Whan Lee, back-warping warps a face image onto the reference image using the corresponding shape information [1]. In our project, we use imwarp, a function provided by MATLAB, to perform any necessary warping to images. The texture information is a matrix of the same dimensions as the input image.

3 Error Back-Projection Procedures

Notation	Definition
L^I	Input low-resolution data
t	Iteration index, $t = 1, 2, \cdots, T$
H_t^R	Reconstructed high-resolution data at iteration t
L_t^R	Low-resolution data simulated by down-sampling reconstructed one at iteration t
D_t^L	Reconstruction error measured by Euclidean distance between input and simulated low-resolution data at iteration t
T_1	Threshold value to determine whether the reconstruction is accurate or not
T_2	Threshold value to determine whether the iteration is convergent or not
L_t^E	Evaluated low-resolution error data by pixel-wise difference between input and simulated low-resolution data at iteration t
H_t^E	Reconstructed high-resolution error of low-resolution error at iteration t
ω_t	Weight for error compensation at iteration t

Figure 1: Notations pertaining to the Error Back-Projection Algorithm

The goal of error back-projection is to iteratively decrease the error of the estimated high-resolution shape and texture of a given low-resolution shape and texture. As observed in Figure 2, this process consists of first obtaining an estimated high-resolution shape or texture by solving least square minimization, a previous step in the experiment pipeline that will explained in section 4. The necessary notation to understand Figure 2 is contained above in Figure 1.

Next, the high-resolution estimate is down-sampled to a low resolution and its Euclidean distance from the original low-resolution shape or texture is measured. If its distance is not below a certain threshold, the error is calculated by taking the difference between the original low-resolution shape or texture and the down-sampled high-resolution estimate.

Afterwards, the error is reconstructed using least square minimization and then added to the previous high-resolution estimate with a empirically derived weight to lessen its error. This process is then repeated until either the reconstruction error is below a certain threshold, the algorithm converges, or the maximum number of allowed iterations is reached.

Because this process relies on using the previous reconstruction error to improve the current reconstruction, the error will always decrease each iteration, resulting in a significant improvement of the high-resolution estimate.

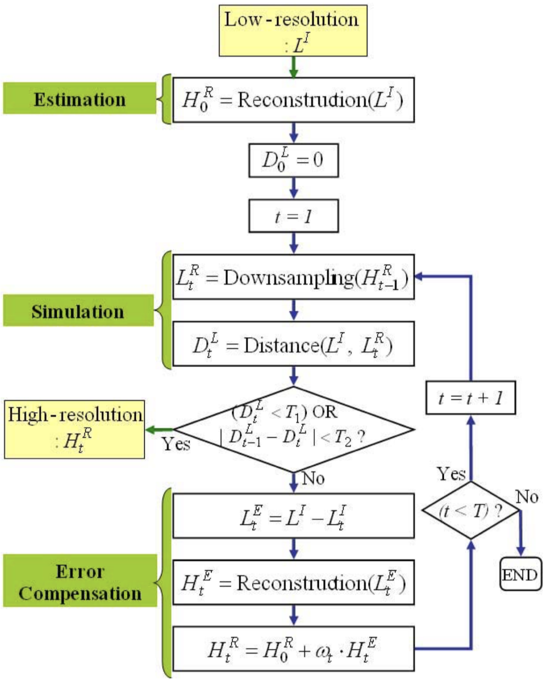
Low-resolution : L^I
Estimation
H_0^R = Reconstruction(L^I)
$D_0^L = 0$
$t = 1$
Simulation
L_t^R = Downsampling(H_{t-1}^R)
D_t^L = Distance(L^I, L_t^R)
High-resolution : H_t^R
Yes
($D_t^L < T_1$) OR
$| D_{t-1}^L - D_t^L | < T_2$?
No
$t = t + 1$
Yes
$(t < T)$?
No
END
$L_t^E = L^I - L_t^I$
Error Compensation
H_t^E = Reconstruction(L_t^E)
$H_t^R = H_0^R + \omega_t \cdot H_t^E$

Figure 2: The Error Back-Projection Algorithm

4 Model Overview

The structure of our model is shown in Figure 3. We will explain each step in the following passage.

Step 1.	Obtain the texture of a low-resolution facial image by backward warping.
Step 2.	(a) Estimate a high-resolution shape from the given low-resolution shape. (b) Improve the high-resolution shape by recursive error back-projection.
Step 3.	(a) Estimate a high-resolution texture from the obtained low-resolution texture obtained at Step 1. (b) Improve the high-resolution texture by recursive error back-projection.
Step 4.	Synthesize a high-resolution facial image by forward warping the estimated texture with the estimated shape.

Figure 3

In our project, we first split the images within our dataset into training and testing images. We then obtain low-resolution pairs for all the images by using the Bicubic interpolation technique to down-sample the images. We then run optical flow algorithm on all low-resolution and high-resolution image pairs of the training set to obtain a pixel-based correspondence, namely the shape information matrix of each image, using the mean face as the reference image. Next, we use back-warping techniques to obtain the texture information of each face. Lastly, we vectorize and then concatenate the shape and texture information to obtain the matrices shown in Figure 4.

$$\begin{aligned} S^+ &= (d_1^x,\ d_1^y,\ \cdots,\ d_L^x,\ d_L^y,\ d_{L+1}^x, d_{L+1}^y\ \cdots,\ d_{L+H}^x,\ d_{L+H}^y)^T \\ T^+ &= (i_1, \cdots, i_L, i_{L+1}, \cdots, i_{L+H})^T \end{aligned}$$

Figure 4

In Figure 4 above, S^+ is composed of the displacement between each pixel and its corresponding pixel in the reference face, and T^+ is composed of the texture intensity of each element in the face. L denotes the number of pixels in the low-resolution image, and H denotes the number of pixels in the high-resolution image. We then apply principal component analysis on both the shape and texture information to obtain the rep-

$$S^+ = \bar{S}^+ + \sum_{p=1}^{M-1} \alpha_p {s_p}^+, \quad T^+ = \bar{T}^+ + \sum_{p=1}^{M-1} \beta_p {t_p}^+ \quad (2)$$

Figure 5

s_p^+ and t_p^+ are the principal component eigenvectors of the shape and texture information. M denotes the number of bases of principal component eigenvectors. Furthermore, $\overline{S}^+$ and $\overline{T}^+$ are the mean shape and textures; meanwhile, α_p are the optimal coefficients for a given image. Lastly, S^+ and T^+ are the shape and texture information of an inputted image.

For the testing set, since only low-resolution faces are available, we want to estimate the high-resolution shape and texture information using this principal component analysis representation. We aim to find a set of optimal α_p which satisfies the following equation in Figure 6.

$$\tilde{S}^+(x_j) = \sum_{p=1}^{M-1} \alpha_p s_p^+(x_j), \ \ j = 1, \cdots, L, \qquad (3)$$

Figure 6

However, there might not be a set of α_p which fit S^+ exactly. Thus, the best we can do is to find a set of α_p to minimize the error of estimation. The problem then becomes one of finding a solution of least square minimization. The equation above can be rewritten as shown in Figure 7.

$$\begin{pmatrix} s_1^+(x_1) & \cdots & s_{M-1}^+(x_1) \\ \vdots & \ddots & \vdots \\ s_1^+(x_L) & \cdots & s_{M-1}^+(x_L) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{M-1} \end{pmatrix} = \begin{pmatrix} \tilde{S}^+(x_1) \\ \vdots \\ \tilde{S}^+(x_L) \end{pmatrix}$$

$$\mathbf{S}^+ \ \vec{\alpha} = \tilde{\mathbf{S}}^+$$

Figure 7

From Figure 7 above, we obtain the following definition shown in Figure 8.

$$\mathbf{S}^+ = \begin{pmatrix} s_1^+(x_1) & \cdots & s_{M-1}^+(x_1) \\ \vdots & \ddots & \vdots \\ s_1^+(x_L) & \cdots & s_{M-1}^+(x_L) \end{pmatrix}$$

$$\vec{\alpha} = (\alpha_1, \cdots, \alpha_{M-1})^T,$$

$$\tilde{\mathbf{S}}^+ = (\tilde{S}^+(x_1), \cdots, \tilde{S}^+(x_L))^T.$$

Figure 8

In our case of facial analysis, we assume that both the columns of S^+ are linearly independent and that $\mathrm{S}^{+^T}\mathrm{S}^+$ is non-singular. The least square solution can be represented in the following form, as shown in Figure 9.

$$\vec{\alpha}^* = (\mathbf{S}^{+^T}\mathbf{S}^+)^{-1}\mathbf{S}^{+^T}\tilde{\mathbf{S}}^+$$

Figure 9

Using the optimal coefficients above, we are able to synthesis high-resolution faces using the following equation shown in Figure 10.

$$S(x_{L+j}) = \bar{S}^+(x_{L+j}) + \sum_{p=1}^{M-1} \alpha_p^* s_p^+(x_{L+j}),\ j = 1, \ldots, H$$

Figure 10

We then use forward warping, which warps the texture information onto input faces according to the shape matrix, to obtain our estimated high-resolution faces.

5 Face Dataset

We use 200 faces from the Max-Planck-Institute Facial Expression Database. The dataset consists of 3D head models recorded by a laser scanner. They all have the same orientation, illumination, similar face expressions, and the background has been completely removed. The dataset is split into 100 faces for training and 100 faces for testing. We also extend our project to the Zhu Face Dataset. This second dataset consists of 177 faces with varying expressions. Like the Max-Planck-Institute Facial Expression Database, the backgrounds are completely removed. However, the ears and scalp are completely removed, leaving only a face rather than a complete head.

6 Results

After running the model on the 100 test faces, some of the results are shown below. In Figures 11 and 13, low-resolution test images are shown. Meanwhile, the high-resolution estimates alongside the original high-resolution test images are shown in Figures 12 and 14. The mean errors for all the images per pixel were found to be: 0.4054 for the x-displacement of the shape information; 0.3249 for the y-displacement of the shape information; 9.0772 for the texture information; and 9.3563 for the image intensities.

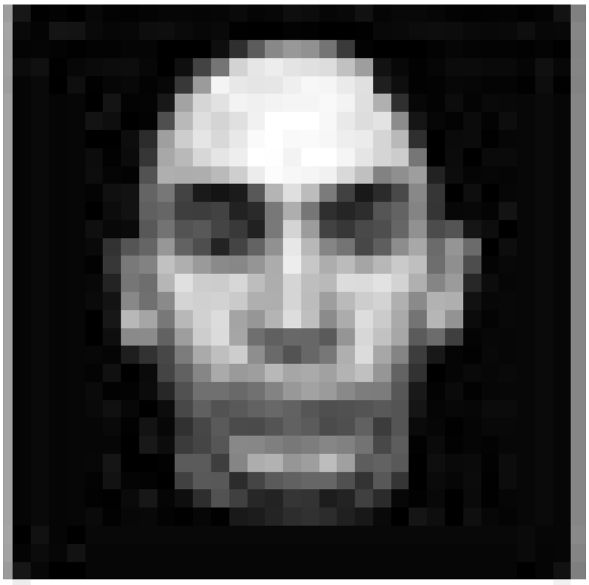

Figure 11

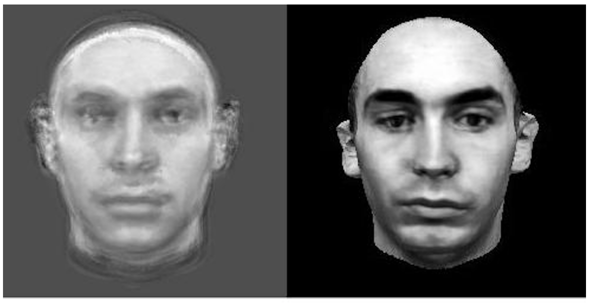

Figure 12

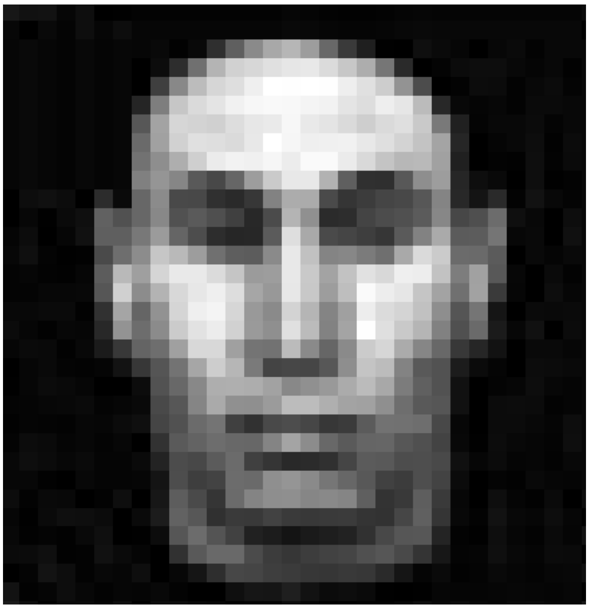

Figure 13

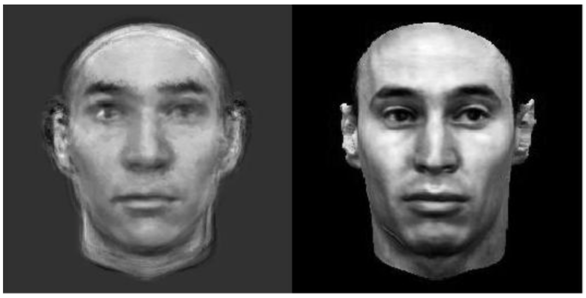

Figure 14

Our model was then retrained on the Zhu Face Dataset. Some of the results are shown below. In Figures 15 and 17, low-resolution test images are shown. Meanwhile, the high-resolution estimates alongside the original high-resolution test images are shown in Figures 16 and 18. The mean errors for all the images per pixel were found to be: 0.7683 for the x-displacement of the shape information; 0.7325 for the y-displacement of the shape information; 10.06225 for the texture information; and 10.8098 for the image intensities.

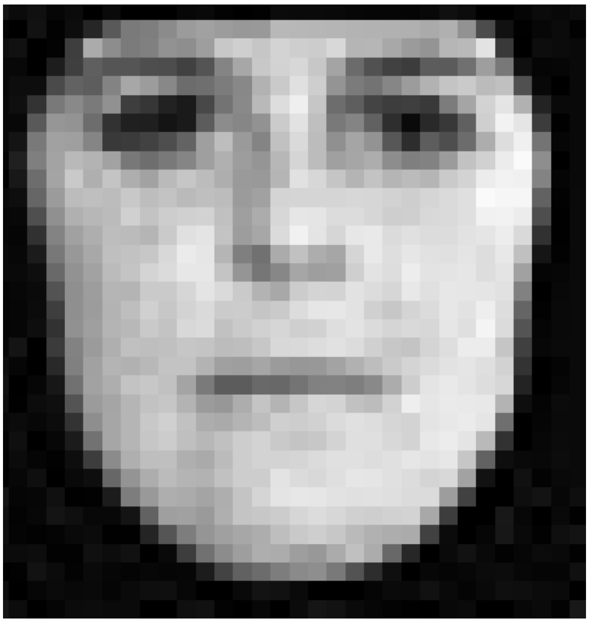

Figure 15

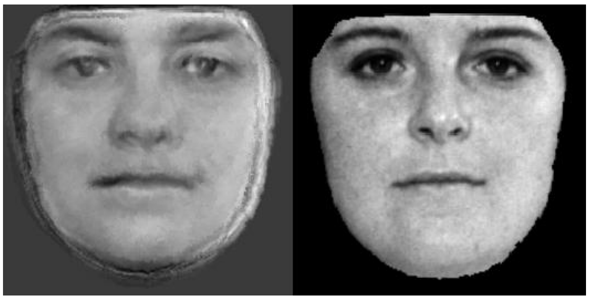

Figure 16

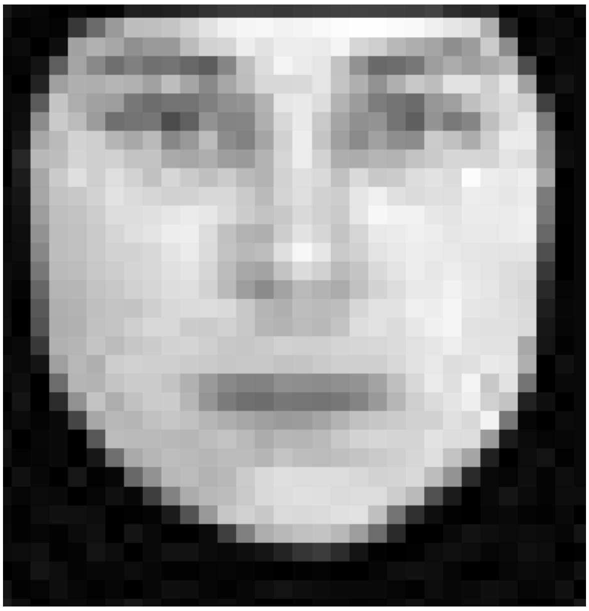

Figure 17

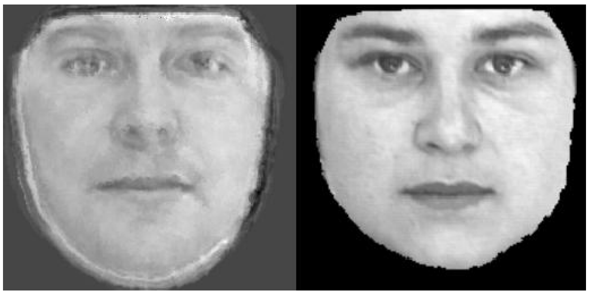

Figure 18

7 Discussion

Analyzing the images, we can see that the high-resolution estimates are satisfactory. The features of the faces are captured with high-definition; however, the faces are somewhat blurry and have some noise. We believe that this is due to our usage of the mean face as the reference image. As the faces are not perfectly aligned, it is inevitable that the mean face has some blurriness.

Analyzing the errors, we can see that the errors for the Zhu Face Dataset are uniformly larger than the errors for the Max-Planck-Institute Facial Expression Database. We reason that this occurs because the faces in this dataset are much larger, causing the shape and texture information between faces to be exaggerated. This makes us harder for the optical flow to obtain a clear shape since the displaceents are much larger. Furthermore, the more jagged edges cut around the face make the mean face much more blurry than the mean face for the MPI Face Database.

More importantly, the faces in the Zhu Face Dataset have a much larger variety of expressions. Thus, depending on the way our dataset was randomly shuffled, some faces may be extremely hard to approximate. For example, if our randomized training dataset did not include a face of someone frowning, then the model would have a difficult time estimating the high-resolution face of someone who is frowning in the original low-resolution image. With such a wide range of expressions, it is difficult to know whether all of them are adequeately captured in the model.

Another cause for the rise in error was that the Zhu Face Dataset was smaller. Thus, the number of bases able to be obtained from principal component analysis were reduced. It is possible that having only 86 bases versus 98 could have caused the slight difference in that the two sets of errors.

8 Conclusion

This project implements a method of enhancing the resolution of facial image from a low-resolution facial image using a recursive error back-projection of example-based learning. Future work includes training this model on a much larger dataset, obtaining better heuristics to choose a reference face, and attempting to utilize different optical flow algorithms to obtain a better correspondence between the low-resolution images and the low-resolution reference face used to train the model. Implementing any of these modifications can lead to improved results in the high-resolution estimates of the low-resolution images.

9 References

[1] J.S. Park, S.-W. Lee, “Resolution enhancement of facial image using an error back-projection of example-based learning”, Proc. 6th IEEE Conf. Automatic Face and Gesture Recognition, pp. 831-836, 2004.
[2] T. Vetter and N. E. Troje, “Separation of texture and shape in images of faces for image coding and synthesis,” Journal of the Optical Society of America A. Vol. 14, No. 9, pp. 2152-2161, 1997.
[3] D. Beymer and T. Poggio, “Image representation for visual learning,” Science 272, pp. 1905-1909, 1996.