FACE SUPER-RESOLUTION FOR LOW-RESOLUTION RGB IMAGES

Atharva Gokhale, Somesh Daga, Somesh Gupta, and Xiaomin Yan

University of Waterloo

ABSTRACT

Face super-resolution presents an attractive research focus due to its wide-ranging applications. The security industry presents an overt application for this field of research, in enhancing the performance of face-recognition for low-quality images obtained through surveillance videos. In this paper, we provide an overview of the face super-resolution space, followed by a comprehensive review of prominent learning and deep-learning based approaches in this field over the past five years. Furthermore, we discuss the limitations of these approaches and highlight challenges that need to be addressed through future works.

***Index Terms*—** Face Super-Resolution, Low Resolution

1. INTRODUCTION

Face super-resolution (SR), otherwise also known as face hallucination, refers to methods of resolving low-resolution images of human faces to higher resolutions that reveal finer, discernible features. It is a domain-specific application of the broader image super-resolution problem, and as such warrants the use of facial structure knowledge to inform the super-resolution process. As highlighted by Chen et al. [1], this area of research garners significant interest due to its anticipated benefits in augmenting the performance of face-related tasks (e.g. face recognition), where performance severely deteriorates with low-resolution inputs.

A general consensus in the field on the classification of low-resolution images are those smaller than 32x32 pixels in size (inferred from [1],[2],[3],[4],[5],[6]). Li et al. [7] draw distinctions between low-quality and low-resolution images, whereby low-quality images may be classified as high-resolution images in certain cases e.g. in the presence of motion blur. Moreover, they argue that real-world images are susceptible to multiple "degradation processes", while a majority of research operates on artificially simulated or "synthetic" images obtained through basic down-sampling of their high-quality counterparts. The super-resolved images resulting from the application of face super-resolution methods are typically upscaled by four or eight times with respect to the low-resolution inputs.

In the following sections, we provide the reader with a comprehensive review of various face super-resolution methods, including the state-of-the-art. Section 2 reviews learning-based methods for face super-resolution applied to face-recognition and 3D face reconstruction tasks. In Section 3, we present deep-learning based techniques that have dominated this field of research in recent years, and discuss the limitations of these works. In addition, we compare and contrast the neural network architectures employed in these methods, and address the datasets and metrics used for training and validation. Finally, in Section 4, we leave the reader with a few concluding remarks and propose directions for future research.

2. LEARNING BASED METHODS

2.1. Background

As noted in [7], face recognition performance is severely inhibited with low quality inputs, and face super-resolution provides an intuitive way to deal with this challenge.

Among learning-based methods, Canonical Correlation Analysis (CCA) has been widely used in recent years. The work of An et. al [8] in the development of a 2D-CCA framework has allowed for this mathematical tool to be applied to face super-resolution, which previously was vastly hindered by 1D-CCA approaches that resulted in largely unrecognizable super-resolved images.

To restore image information in high frequency modes and deal with the dimensional differences between HR and LR images ([9]), another learning-based method named the "Double Layer Coupled Locality Preserving Method" [10], was proposed by Chen et al. The crux of this method was to find the relationship between HR images and LR images and calculate a coupled mapping matrix to measure similarity.

2.2. Metrics

In this section we review metrics used by various learning-based methods.

An et al. [8] argued that despite lower PSNR values, they produced the more visually appealing results in their work as compared with previous works. And so, they proposed the use of the alternative measures of Distortion Measurement (DM) and SVD which had previously been developed for human perception models.

Moreover, a number of metrics exist that are particular to their application. For example, reconstruction error is a widely used metric in the problem of 3D face reconstruction ([11]), while recognition rate is typically used to evaluate face recognition performance in both 2D and 3D domains.

In the calculation process, Chen et al. [10] proposed the concept — similarity discrimination. The experiment mainly considers two aspects, the first is the relationship of the dimension of the coupled space and the very low-resolution image recognition rate, the second is the relationship between the number of training samples and the very low-resolution image recognition rate.

2.3. Methods

Below, we present some important works in learning-based approaches to face super-resolution as applied to face recognition and reconstruction.

An et al.[8] formulated 2D-CCA in such a way that it takes two sets of images without the need to vectorize each set of images. The method of [8] consists of two steps that preserve the intrinsic 2D spatial structure of face images in the super-resolution process. In the first step, the HR face is reconstructed, but as the reconstructed face is not rich in facial details, the model subsequently applies a high-frequency detail mask to the reconstructed faces in a second step.

Aouada et al. [11] proposed another method which consists of three main steps; pre-processing of raw data, feature extraction and matching. In the pre-processing of the raw data, first the face region and nose tip is detected using the method of Viola-Jones [12], as it is a computationally efficient algorithm. Features are extracted in the form of spherical curves, after performing a super-resolution step, by intersecting facial surfaces with a sphere. Lastly, in the matching step, extracted features are compared with features stored in a database to find the closest facial match.

The DLCLPM method [10] initially deals with the similarity discrimination. The most primitive distance formula for coupled mapping is mapping the HR images and the LR images into a unified potential coupled space. The distance measure of feature sets are made in d-dimensional space [9]. The DLCLPM method optimizes the objective function in the coupled spatial mapping and fuse two coupled spaces into one equation [10]. Computational process yields better results for minimum value of both the variables, also the global objective function will obtain minimum value. This problem can be transformed into a standard eigenvalue solving problem.

Concerning the similarity, the mapping matrix between HR image and LR image can be learned in the first layer of coupled space. The euclidean distance between the test sample and the training sample is calculated to obtain a local similarity measure [10]. Considering the images belonging to different scales in double layer coupled space with different feature information, the similarity measure can add the weight term parameters which can balance the scales of the measured distance in double layer coupled space. When the value of the measured distance is minimum, the test sample and the training sample can be paired respectively into the same class.

2.4. Discussion

The model of An et al.[8] is compared with four state-of-the-art learning-based models using the same training sets. In comparison, they claimed to have produced the most visually pleasing results. The results obtained from [8] indicate that both part based and detail compensation improved the quality of the output image. The model is very efficient due to its low computational complexity and also improves the face recognition performance to more than 30 percent using a hybrid data set compared to previous works.

For 3D face recognition, the method of Berreti et al.[13] achieved a recognition rate of 50 percent on the Super Faces dataset. Aouada et al. [11] argued that depth reconstruction artifacts were introduced due to which the recognition rate of the system was reduced. [11] reformulated [13], in which the latter was modified to use surface reconstruction and a deblurring step was added, resulting in a 30 percent increase in recognition rate on the same dataset.

In the experiment for DLCLPM[10] the data is preprocessed using the traditional PCA method to reduce the dimension of original data, and use random extraction method to select samples[9]. The researchers use three face datasets, CMU PIE, Yale and the FERET. The resolution of the training, intermediate and input image is 32 x 18, 16 x 14, and 8 x 7 respectively. The results show that double layer coupled mappings method gets better recognition rate and it also improves the accuracy in all three datasets. On comparing DLCLPM with the previous works such as CM method, CLPM method and LGCM method[10], DLCLPM proves to be the most efficient.

2.5. Limitations

As pointed out in [8], after reconstructing the face, the dc gain in the frequency map was high relative to the higher frequency components that represented the finer features. This was evident from the blurriness of the output images.

Furthermore, we note that the [8] utilized mostly frontal images where face alignment was not an issue. Hence, we hypothesize that the results of this work will not translate well to ill-aligned faces unless a face alignment operation is performed to the input images.

The LPP method used in coupled mappings can only guarantee that the locally distributed samples won't be changed in the feature space, but, it doesn't hold true for the samples that are far away in the original space. Due to the influence of noise, expression and posture, two samples of the same person may be spread out in the original sample space as com-

pared to the samples of different people. This condition is unreasonable in the low-dimensional feature space.

DLCLPM is an extension of LPP, which guarantees the consistency of the selected samples in low and high dimensional spaces [14]. The most import idea is to figure out the minimum similarity discrimination. We have to make sure that the selected samples in the low dimensional feature space are consistent with selected in the original high dimensional image space [9].

3. DEEP-LEARNING BASED METHODS

3.1. Background

Deep-learning methods constitute the majority of recent works in face super-resolution. The incorporation of facial priors in the training process have been largely made possible by a number of publicly accessible datasets, complete with high-quality ground truth images with annotated facial landmarks and/or heatmaps, and state-of-the-art research in the field of human pose estimation.

In addition, Generative Adversarial Networks (GANs) have been adapted into a number of recent works ([1],[2], [4], [5], [6]), to produce photo-realistic, super-resolved images. GANs are fundamentally comprised of two competing neural networks, termed the generators and discriminators. Applied to the face super-resolution problem, the discriminator attempts to differentiate between high-quality facial images and super-resolved ones generated by the super-resolution (or generator) network. A corresponding loss termed the "Adversarial Loss" is employed to train these networks.

Deep-learning based super-resolution methods are often classified as multi-stage or end-to-end. The former requires the use of multiple separately trained neural networks to generate the final, super-resolved image. Multi-stage methods used in prior research in this field have been largely replaced by end-to-end learning techniques as in [1], [2],[4],[5]. The latter generally use intermediate supervision, whereby losses are computed for outputs at multiple probe points in a network and added together to evaluate a total combined loss, which is subsequently used to adjust weights across the entire network. [1] argues that the act of training weights across the entire network in unison leads to better results, as all weights factor in the quality of the final, super-resolved image during training.

3.2. Metrics

[1],[2],[6] echo similar sentiments on the incongruity of established quantitative measures of PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) with qualitative results, in evaluating super-resolution performance. They argue that defining losses to optimize such metrics, bias super-resolution methods to generate overall smooth images as opposed to sharper images revealing finer features, thereby contradicting the aim of the face super-resolution challenge.

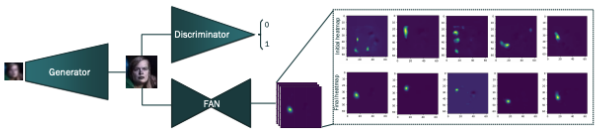

Fig. 1. Super-FAN network architecture (Obtained from [2])

As a result, [1] and [2] have separately proposed face alignment/landmark detection as alternative measures that reconcile quantitative and qualitative results. The rationale is that the performance of face landmark localization algorithms is based on the perceptual quality of the inputs i.e. the super-resolved images. The usage of the proposed measures are in-line with the definition of perceptual losses (discussed in 3.4) used in recent studies to produce state-of-the-art results.

3.3. Methods

In this section, we provide an overview of the prominent deep-learning works in face super-resolution in recent years. These methods operate on 16x16 low-resolution images and super-resolve them by a scale of four or eight times. State-of-the-art results in face super-resolution are jointly represented by [1] and [2] with [6] as the baseline. Their network architectures are depicted in figures 2 and 1 respectively.

Bulat et al. [2] proposed a residual-based super-resolution network, trained in an end-to-end fashion, in conjunction with a discriminator and face alignment network (FAN) for face landmark localization. Using landmark localization accuracy as their primary measure, benchmarked via the AUC method with a threshold of 10%, they demonstrated state-of-the-art results across a range of facial poses. Of note, the upper bound for localization accuracy was established using a pre-trained face alignment network on the high-quality images obtained from the utilized datasets.

Chen et al. [1] proposed a similar network to [2], with the primary difference being that face landmark localization was performed on intermediate upsampled images, as opposed to their final super-resolved counterparts. They presented their results via an Intersection-over-Union (IoU) approach whereby the overlap between the facial heatmaps generated by their networks and those directly obtained from the utilized datasets were compared. Moreover, results were provided for both their primary network, FSRNet, and a GAN variant of their network (known as FSRGAN).

3.4. Discussion

We reserve this section for the discussion of materials that have have enabled the aforementioned works.

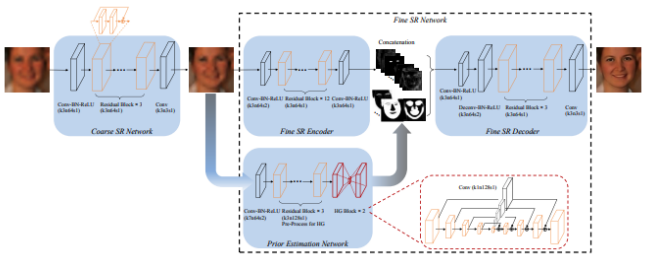

Fig. 2. FSRNet network architecture (Obtained from [1])

Notably, both the FSRNet [1] and Super-FAN [2] networks utilize the "hourglass" modules of [15] for their landmark estimation networks. This work has been adapted into the face super-resolution space, from the more general task of localizing human joints in images. An hourglass module is represented by a neural network, that extracts features at different scales of an image, such that global and local context-awareness is used to inform the localization. The state-of-the-art results exhibited by this work in the area of human pose estimation were largely touted for their increased accuracy in tracking difficult joints like ankles or knees, as opposed to localizing features in faces.

In addition, [1],[2],[4],[6] incorporate the use of "perceptual loss" functions to super-resolve faces. [16] demonstrated qualitative improvements in edge and detail reconstruction performance of the super-resolution network of [17] through the introduction of such loss functions. They represented these losses by a pixel-wise loss of feature representations obtained through the use of pre-trained, image classification neural networks. The aforementioned works utilize the concept of perceptual losses through a number of different feature representations such as face landmarks and local/global heatmaps.

The task of comparing results across studies is complicated by the shifting measures of super-resolution performance. Even though a general consensus is trending towards the use of face localization measures, the implementation and feature representations that quantify these measures is not established. Moreover, the variation in datasets used for training and validation introduce further issues. As argued by [2], numerous datasets like Celeb-A contain near-frontal poses of faces while other datasets may contain an even balance of frontal, intermediate and extreme poses. Furthermore, they contend that few papers address the performance of super-resolution on real-world images through datasets like WiderFace.

3.5. Limitations

In this section, we analyze some limitations of the deep-learning methods presented previously.

Firstly, we draw the reader's attention to the neural network architectures depicted in figures 1 and 2. It can be observed that the network of [2] does not require the discriminator or face alignment network to be executed during inference, since the super-resolved image is generated prior to the execution of those networks. Contrasting this with the network of [1], we see that the super-resolution process needs to be informed by the so-called prior estimation network, and hence the entire network needs to be run during training and inference. This introduces some consequences with respect to the trade-off between face landmark localization accuracy and time complexity for [1]. The hourglass modules of [15] used by both the aforementioned methods for landmark localization, improve localization accuracy as more of these modules are stacked end-to-end. However, this increases the inference time for the localization network. Therefore, there are repercussions in the inference time complexity for the method of [1] which do not impact [2]. As a result, we emphasize that the structure of the entire network plays an important role in the time-complexity of super-resolution performance, even though the networks are comprised of similar components.

Secondly, the use of the hourglass modules of [15] impose further limitations on the super-resolution networks that utilize them. [15] cites a significant likelihood of failure in localizing features when multiple people are present in an image. Common modes of failure include partial localization of features across all individuals in the image. In real images, it can be expected that there might be a number of people in an image, and in some scenarios, the task of image segmentation to remove unwanted persons may not be easy. Hence, super-resolution methods need additional mechanisms to address the existence of multiple people in an image.

Lastly, [1]-[5] use datasets such as CelebA, Menpo, Vggface2 and Helen which mostly include front facing images that are used for training various neural networks. As a result these networks are severely impacted by inputs with extreme poses and occlusions.

4. CONCLUSION

Deep-learning methods utilizing facial priors have largely superseded learning-based methods. State-of-the-art deep-learning approaches have demonstrated strong results on a number of public datasets. However, the use of simulated low-resolution images do not necessarily translate well to real-world images that may contain low-quality artifacts other than those arising from low-spatial resolutions of cameras. Moreover, methods in this area of research continue to be plagued by a number of challenges like face alignment, occluded faces and the presence of multiple people in an image. Hence, we propose these as directions for future research. Considering [18]

5. REFERENCES

[1] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2492–2501, 2018.

[2] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 109–117, 2018.

[3] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1698–1706, Oct 2017.

[4] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Computer Vision – ECCV*, pp. 219–235, 2018.

[5] B. Dogan, S. Gu, and R. Timofte, "Exemplar guided face image super-resolution without facial landmarks," *CoRR*, vol. abs/1906.07078, 2019.

[6] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 105–114, 2017.

[7] P. Li, P. Flynn, L. Prieto, and D. Mery, "Face recognition in low quality images: A survey." arXiv preprint arXiv:1805.11519, 2018.

[8] L. An and B. Bhanu, "Face image super-resolution using 2d cca," *Elsevier Signal Processing*, vol. 103, pp. 184–194, 2014.

[9] X. Zhang, J. Jiang, J. Li, and S. Peng, "Manifold learning-based sample selection method for facial image super-resolution," *Society of Photo-Optical Instrumentation Engineers (SPIE)*, vol. 51, p. 7003, April 2012.

[10] H. Chen, Y. Zhang, and J. Pei, "Double layer coupled locality preserving mappings for very low-resolution face recognition," pp. 63–67, March 2019.

[11] D. Aouada, K. A. Ismaeil, K. K. Idris, and B. Ottersten, "Surface up-sr for an improved face recognition using low resolution depth cameras," in *11th IEEE Inter. Conf. Adv. Video Sign. Surv. (AVSS)*, pp. 107–112, 2014.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," vol. 1, pp. I–I, February 2001.

[13] S. Berretti, A. Del Bimbo, and P. Pala, "Superfaces: A super-resolution model for 3d faces," in *Computer Vision – ECCV 2012. Workshops and Demonstrations*, pp. 73–82, 2012.

[14] W. Dong, l. Zhang, and G. Shi, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Proc.*, vol. 20, pp. 1838–1857, 2011.

[15] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision - ECCV*, vol. 9912, pp. 483–499, 2016.

[16] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision - ECCV*, vol. 9906, pp. 694–711, 2016.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 38, pp. 295–307, January 2016.

[18] S. Ruder, "An overview of gradient descent optimization algorithms.," 2016. cite arxiv:1609.04747Comment: Added derivations of AdaMax and Nadam.