MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Super-resolution of Very Low-Resolution Faces from Videos

Ataer-Cansizoglu, Esra; Jones, Michael J.

TR2018-140 September 26, 2018

Abstract

Faces appear in low-resolution video sequences in various domains such as surveillance. The information accumulated over multiple frames can help super-resolution for high magnification factors. We present a method to super-resolve a face image using the consecutive frames of the face in the same sequence. Our method is based on a novel multi-input-single-output framework with a Siamese deep network architecture that fuses multiple frames into a single face image. Contrary to existing work on video super-resolution, it is model free and does not depend on facial landmark detection that might be difficult to handle for very low-resolution faces. The experiments show that the use of multiple frames as input improves the performance compared to single-inputsingle-output systems.

British Machine Vision Conference (BMVC)

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Super-resolution of Very Low-Resolution Faces from Videos

Esra Ataer-Cansizoglu
cansizoglu@merl.com
Michael Jones
mjones@merl.com

Mitsubishi Electric Research Labs
(MERL)
Cambridge, MA, USA

Abstract

Faces appear in low-resolution video sequences in various domains such as surveillance. The information accumulated over multiple frames can help super-resolution for high magnification factors. We present a method to super-resolve a face image using the consecutive frames of the face in the same sequence. Our method is based on a novel multi-input-single-output framework with a Siamese deep network architecture that fuses multiple frames into a single face image. Contrary to existing work on video super-resolution, it is model free and does not depend on facial landmark detection that might be difficult to handle for very low-resolution faces. The experiments show that the use of multiple frames as input improves the performance compared to single-input-single-output systems.

1 Introduction

Face images appear in various platforms and are vital for many applications ranging from forensics to health monitoring. In most cases, these images are low-resolution, making face identification difficult. Therefore, upsampling low-resolution face images is crucial. In this paper, we present a method to generate a super-resolved face image from a given low-resolution (LR) face video sequence.

The use of multiple frames for super-resolution is well studied. However, previous work mostly focuses on images of general scenes and the magnification factor does not go beyond $4\times$. Moreover, most of these algorithms rely on motion estimation between frames, which is hard when the target object in the image is very low-resolution. There are a few studies addressing the problem of face super-resolution using videos [8, 9, 28, 29]. These methods are either model-based [8, 9] or registration-based [28, 29]. Therefore, the magnification factor is small since the model fitting and registration stages highly depend on landmark detection or texture, respectively. In order to address these limitations, we present a model-free multi-frame face super-resolution technique designed for very low-resolution face images. The LR faces we target in this work are tiny with a facial area of around 6×6 pixels that are difficult to identify even by humans. Our method uses a deep network architecture consisting of two subnetworks. The first subnetwork is trained to super-resolve each frame independently, while the second subnetwork generates weights to fuse super-resolved images of all frames.

Thus, the fusion network uses the information accumulated over multiple images in order to generate a single super-resolved image.

Our main contributions are three fold: (i) We present a novel method for multi-frame super-resolution of very tiny face images with a magnification factor of 8 times. (ii) The presented method is model free and requires only rough alignment instead of pixelwise registration such as many previous multi-frame super-resolution algorithms depend on. (iii) Our technique involves a novel multi-input-single-output (MISO) framework with a Siamese deep network architecture that fuses multiple frames into a single face image. Instead of relying on a prior motion estimation stage that is challenging for very LR face images, our fusion network implicitly uses the motion among frames in order to generate the final image.

1.1 Related Work

The use of multiple frames to super-resolve a single image is well studied in the literature [18]. There are a wide range of techniques including, iterative [6, 10], direct [5, 21] and probabilistic [1, 20, 32] approaches for performing multi-input-single-output (MISO) super-resolution. The early methods [6, 10] focused on iterative back projection in order to estimate the forward imaging model given low-resolution sequences. Most of these methods deal with frames with known transformations between them making it hard to apply to real scenarios. In the category of direct methods, registration among frames is estimated and utilized for producing a single output [5, 21]. Since registration plays an important role in the super-resolution process, these methods are not capable of handling high magnification factors, especially if the target object appears very small in the input image. Probabilistic approaches [1, 20, 32] utilize different regularization terms depending on the sufficiency of the given low-resolution images and the magnification factor. Designing the noise model and tuning regularization parameters are challenging due to the fact that the problem is ill-posed.

With the rise in deep learning algorithms, recently there have been attempts to improve super-resolution using neural networks. Some methods considered the registration among frames. Liao *et al*. [14] presented a method that first generates super-resolution draft ensemble by using an existing motion estimation method and then reconstructing the final result through a CNN deconvolving the draft ensemble. Tao *et al*. [22] introduces a subpixel motion compensation layer in order to handle inter-frame motion. Ma *et al*. [17] presents a motion de-blurring method based on an expectation-maximization framework for estimating least blur and reconstruction parameters. Kappeler *et al*. [12] experiments with various CNN architectures for video super-resolution by incorporating neighboring frames alignment. All of these methods are for general video SR and they only aim for at most $4\times$ magnification.

There has been a tremendous amount of work for single-image super-resolution of faces [2, 4, 13, 15, 16, 24]. Recently, a few papers have used convolutional neural networks (CNNs) or generative adversarial networks (GANs) for face-specific super-resolution. These methods better handle uncontrolled input faces with variations in lighting, pose and expression and only rough alignment. Zhou et al. [34]'s bi-channel approach used a CNN to extract features from the low-resolution input face and then mapped these using fully connected network layers to an intermediate upsampled face image, which is linearly combined with a bicubicly interpolated upsampling of the input face image to create a final high-resolution output face. In other work, Yu and Porikli [30, 31] used GANs for 8 times super-resolution of face images. Cao et al. [3] presented a deep reinforcement learning approach that sequentially discovers attended patches followed by facial part enhancement. Zhu et al. [35] proposed a bi-network architecture to solve super-resolution in a cascaded way. Each of these methods

uses a single image of the person to produce the final super-resolved image. Therefore, they do not make use of the information coming from multiple frames.

Multiple frames for face super-resolution have been used in early studies in the field. However, the datasets were generated with simple motion translation with alignment and the faces were captured under controlled environment. Baker *et al*. [1] followed a Bayesian formulation of the problem and presented a solution based on reconstruction and recognition losses. Their technique makes use of multiple frames that are generated by simple shifting transformations. Recently, Huber *et al*. [8, 9] introduced a video-based super-resolution algorithm that makes use of the Surrey face model. The technique fuses intensities coming from all frames in a 2D isomap representation of the face-model. It highly depends on landmark detection, thus it is not possible to apply to very low-resolution faces such as the ones used in this study. Two other related studies [28, 29] focus on the problem of face registration for MISO face super-resolution. Yu *et al*. [29] presented a method to perform global transformation and local deformation for $3\times$ upsampling of faces with expression changes. Similarly, Yoshida *et al*. [28] employs free-form deformation method to warp all frames into a reference frame followed by fusion of warped images. In both studies, the facial area in input low-resolution images is large (i.e. greater than 30×20 pixels), which enables an accurate registration.

2 Method

2.1 Problem Statement and notation

Given a sequence of N low-resolution $W \times H$ face images, $S_L = \{\mathbf{I}_L^1, \mathbf{I}_L^2, \cdots, \mathbf{I}_L^N\}$, our goal is to predict a corresponding high-resolution (HR) face image $\hat{\mathbf{I}}$ with a size $dW \times dH$ where $d \in \mathbb{N}$ is a non-zero magnification factor. Each frame in the sequence is a LR version of the face in different poses and expressions. Thus

$$\mathbf{I}_L^k = \mathbf{g}(\mathbf{f}_k(\hat{\mathbf{I}})) = \mathbf{D}\mathbf{H}\mathbf{f}_k(\hat{\mathbf{I}}) \tag{1}$$

where $\mathbf{f}_k : \mathbb{R}^{dW \times dH} \rightarrow \mathbb{R}^{dW \times dH}$ denotes a deformation function that takes a face and transforms it to another pose and expression and $\mathbf{g} : \mathbb{R}^{dW \times dH} \rightarrow \mathbb{R}^{W \times H}$ is the function representing the LR image capturing process, which is composed of a blurring matrix $\mathbf{H}$ that represents point spread function and a downsampling matrix $\mathbf{D}$.

2.2 Framework

During training, we are also provided corresponding HR image sequence $S_H = \{\hat{\mathbf{I}}_H^1, \hat{\mathbf{I}}_H^2, \cdots, \hat{\mathbf{I}}_H^N\}$. We are interested in finding inverse of both $\mathbf{f}_k$ and $\mathbf{g}$ functions in order to obtain an estimate for $\hat{\mathbf{I}}$. Considering the sequential nature of the process, we follow a two-step approach. At the first step, we learn a generator to estimate HR version $\hat{\mathbf{I}}_H^k$ of each image in the sequence. Second, we estimate $\hat{\mathbf{I}}$ from super-resolved images by learning a fusion function that represents $\mathbf{f}_1^{-1}, \mathbf{f}_2^{-1}, \cdots, \mathbf{f}_N^{-1}$.

Let $\mathbf{G} : \mathbb{R}^{W \times H} \rightarrow \mathbb{R}^{dW \times dH}$ be a generator function that creates super-resolved images given LR images. We learn $\mathbf{G}$ by minimizing the following cost function:

$$\mathfrak{L}_{gen} = \frac{1}{TN} \sum_{<S_L, S_H>} \sum_{i}^{N} \mathfrak{D}(\mathbf{G}(\mathbf{I}_L^i), \hat{\mathbf{I}}_H^i) \tag{2}$$

where T is the number of sequences in the training set and $\mathfrak{D}$ is a distance measure between two images.

The fusion function is parametrized by a sequence of weight maps, which produces a weighted sum of all the super-resolved images

$$\mathbf{F}(S_L,\mathbf{G}) = \sum_{i}^{N} \mathbf{F}_i(\mathbf{I}_L^1,\cdots,\mathbf{I}_L^N,\mathbf{G}(\mathbf{I}_L^1),\cdots,\mathbf{G}(\mathbf{I}_L^N)) \otimes \mathbf{G}(\mathbf{I}_L^i) \tag{3}$$

where $\mathbf{F}_i : (N \times \mathbb{R}^{W\times H}) + (N \times \mathbb{R}^{dW\times dH}) \rightarrow \mathbb{R}^{dW\times dH}$ denotes a function that returns pixel weights given a sequence of LR and corresponding super-resolved images and $\otimes$ indicates element wise multiplication between matrices. $\mathbf{F}$ is learned by minimizing the fusion cost

$$\mathfrak{L}_{fus} = \frac{1}{T} \sum_{<S_L,S_H>} \mathfrak{D}(\mathbf{F}(S_L,\mathbf{G}),\hat{\mathbf{I}}). \tag{4}$$

$\hat{\mathbf{I}}$ is taken as HR version of the K^{th} frame from the sequence.

We employ two types of distance measures in training: L2 distance and structural similarity (SSIM) distance. L2 distance between two images $\mathbf{X},\mathbf{Y} \in \mathbb{R}^{W\times H}$ is defined as

$$\mathfrak{D}_{L2}(\mathbf{X},\mathbf{Y}) = \|\mathbf{X}-\mathbf{Y}\|_F \tag{5}$$

where $\|\cdot\|_F$ indicates Frobenius norm. SSIM is a popular measure that accounts for humans perception of image quality [33]. On a local patch of two images $\mathbf{X}$ and $\mathbf{Y}$ around pixel $\mathbf{p}$, SSIM is computed as $SSIM(\mathbf{X},\mathbf{Y},\mathbf{p}) = \frac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1}\frac{2\sigma_{xy}+c_2}{\sigma_x^2+\sigma_y^2+c_2}$,where μ_x,μ_y and σ_x,σ_y denote the mean and variance, respectively, of the intensities in the local patch around pixel $\mathbf{p}$ in $\mathbf{X}$ and $\mathbf{Y}$ and σ_{xy} is the covariance of intensities in the two local patches. c_1 and c_2 are constant factors to stabilize the division with weak denominator. We divide the image into $h\times h$ grids and employ the mean of SSIM over all patches to come up with the SSIM distance between two images

$$\mathfrak{D}_{ssim}(\mathbf{X},\mathbf{Y}) = \frac{1}{T}\sum_{k=1}^{T}\sum_{\mathbf{p}}\Big[1-SSIM(\mathbf{X}^k,\mathbf{Y}^k,\mathbf{p})\Big] \tag{6}$$

where T is the total number of patches and $\mathbf{X}^k$ and $\mathbf{Y}^k$ is the kth corresponding patch pair. Compared with reconstruction loss, minimizing SSIM loss helps recover the information at high frequency visually. Consequently, we utilize a weighted sum of the two distance measures in training our algorithm

$$\mathfrak{D}(\mathbf{X},\mathbf{Y}) = \mathfrak{D}_{L2}(\mathbf{X},\mathbf{Y}) + \alpha\mathfrak{D}_{ssim}(\mathbf{X},\mathbf{Y}). \tag{7}$$

2.3 Network Architecture

Overview of our system is seen in Figure 1. Each LR image in the input sequence is first given to the generator network to obtain an estimate of their HR version. LR images are upsampled to the HR image size with bicubic interpolation[1] and input to the fusion network along with the super-resolved images. The fusion network generates weight maps based on the temporal information in the sequence.

[1] Any other upsampling technique can be used here.

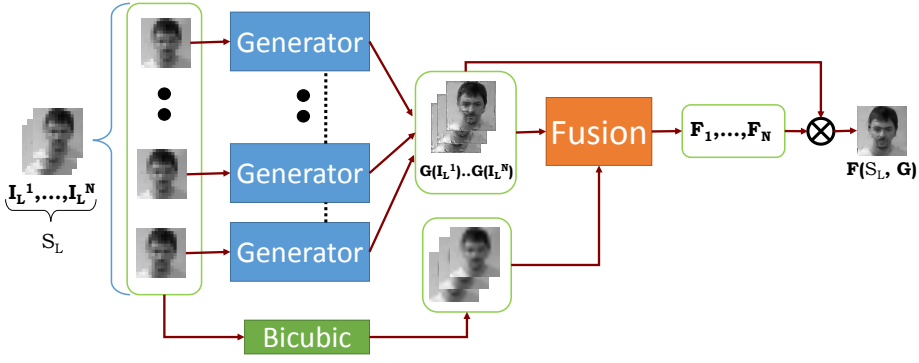

Figure 1: System overview. Dashed lines indicate weight sharing between subnetworks.

Network	Layer Index (Depth)	Type	Kernel Size	Stride	Pad	Output Channels	Learning Rate
Generator	1	Conv	3	1	1	512	10^{-4}
Generator	2	Deconv	2	2	0	512	0
Generator	3	Conv	3	1	1	256	10^{-4}
Generator	4	Deconv	2	2	0	256	0
Generator	5	Conv	5	1	2	128	10^{-4}
Generator	6	Deconv	2	2	0	128	0
Generator	7	Conv	5	1	2	64	10^{-4}
Generator	8	Conv	5	1	2	1	10^{-4}
Fusion	1-5	Conv	3	1	1	64	10^{-5}
Fusion	6	Conv	3	1	1	1	10^{-6}

Table 1: Network architecture

For generator we use a slightly modified version of [30] trained on gray-scale images. For fusion network we use a similar architecture with the gated network of [35]. The details of our architecture is given in Table 1. Each convolution layer is followed by RELU except the last layers. The deconvolution layers perform a simple nearest neighbor upsampling of the data and their weights are kept fixed during training.

3 Experiments

We carried out experiments in controlled and uncontrolled environments on three datasets. We first performed experiments on FRGC dataset [19], which consists of frontal face images captured under a controlled environment. We generated a video sequence for each image in the dataset by simulating a similarity transform. In order to analyze the performance of the method on real-life scenarios, we did experiments on ChokePoint [26] and YouTube Faces [25] datasets, which contains videos of faces in-the-wild.

3.1 Dataset and Experimental Setup

FRGC: The FRGC dataset contains frontal face images taken in a studio setting under two lighting conditions with only two facial expressions (smiling and neutral). We generated training and test splits, where we kept the identities in each set disjoint. The training set consisted of $20,000$ images from 409 subjects and the test set consisted of $2,149$ images from 142 subjects. In order to generate a sequence of face images, for each image in the training or testing set, we randomly generated different similarity transforms and applied them to the HR image, followed by LR image generation.

ChokePoint: This dataset consists of face videos of subjects captured, when they are walking through a portal in a natural way. We used the fourth sequence of the images in P1E, P1L and P2E portals as test set, and the rest as training set, achieving a diverse set of training and test images for enter-exit and indoor- outdoor scenarios. We discarded the frames with smaller facial region than our experimental HR setting. As a result, we ended up having $8,272$ training image sequences and $1,884$ test sequences.

YouTube Faces: This dataset contains $3,245$ videos of $1,595$ people downloaded from YouTube. It comes with 10 folds separated with disjoint identities among them. We used the first fold as the test set and the rest as training set. As a result, we had $46,693$ test image sequences and $402,003$ training image sequences.

Data preparation: HR images had 128×128 pixel resolution, where the faces occupied approximately 50×50 pixels. We generated low-resolution images with 8 times downsampling by following the approach in [27]. As a result, the LR images contained a facial area of approximately 6×6 pixels. We used $N = 8$ consecutive frames in forming our MISO datasets. The images in the sequence were roughly aligned following a similarity transform based on location of eye and mouth centers. ChokePoint dataset already provides facial landmark points as ground truth. For FRGC and YouTube Faces datasets, we employed landmark detection method in [23]. In practice, a low-resolution face detector would be sufficient for rough alignment. We set $K = 8$ and superresolved the last frame in each sequence.

Algorithm Details and Experimental Setup: We trained two different networks depending on the distance measure used. In other words, we either used only L2 distance by setting $\alpha = 0$ or used both L2 and SSIM distances as a joint distance measure. The value of α is selected with a greedy search procedure as $\alpha = 1000$. We super-resolved test image sequences using our multi-input-single-output (MISO) super-resolution framework. Since the first part of our network architecture (i.e. generator network) is considered single-input-single-output (SISO), we also performed experiments by inputting the test images one by one to the generator network. Our goal is to better understand the gain in using video inputs as opposed to single images.

The code is implemented using Caffe deep learning framework [11]. Optimization was performed using the RMSProp algorithm [7]. The training took 3 days on a NVIDIA GeForce GTX TITAN X GPU with Maxwell architecture and 12GB memory. Average running time for super-resolution of a sequence is 0.12 second. We used a patch size of $h = 8$ for SSIM loss.

3.2 Quantitative Results

We evaluated the performance of our algorithm using peak signal-to-noise-ratio (PSNR) and structural similarity (SSIM) measures as reported in Table 2. As can be seen using only L2 distance can yield larger PSNR values, but the use of SSIM loss improves SSIM scores as

Method	FRGC		ChokePoint		YouTube Faces	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	23.812	0.606	26.034	0.660	23.591	0.598
VSRnet [12]	22.997	0.565	26.778	0.693	23.907	0.620
SPMC [22]	21.388	0.555	23.083	0.648	21.189	0.539
SISO (only L2 distance)	26.810	0.808	27.353	0.746	24.676	0.682
SISO (joint distance)	26.767	0.809	27.494	0.751	24.791	0.687
MISO (only L2 distance)	26.892	0.811	27.663	0.761	25.088	0.700
MISO (joint distance)	**26.903**	**0.813**	**27.726**	**0.764**	**25.108**	**0.701**

Table 2: PSNR and SSIM values for the experimental results on all datasets.

expected. The improvement from a SISO architecture to MISO architecture is more obvious on ChokePoint dataset. This might be due to the fact that ChokePoint contains subjects that walk in a certain trajectory yielding certain motion patterns in the sequences. Hence, fusion network benefits from the motion pattern in order to generate more accurate super-resolution results.

We also compared our algorithm with two recent existing video super-resolution methods: VSRnet by Kappeler *et al*. [12] and SPMC by Tao *et al*. [22]. Both methods are designed for general scenes with at most $4\times$ magnification. Thus, we applied the methods with $4\times$ upsampling followed by $2\times$ upsampling. We used the provided models by the authors to test on the face datasets. The sequence size is 5 and 3 for VSRnet and SPMC respectively. Thus, the final single image is produced using 9 frames for VSRnet and 5 frames for SPMC method. Since these methods depend heavily on a preprocessing stage for motion estimation, they suffer from the small facial region seen in the LR images. On the contrary our algorithm is more adaptable, where motion information is learned in the fusion network implicitly. As can be seen from the Table 2, our method outperforms VSRnet and SPMC based on PSNR and SSIM measures.

3.3 Qualitative Results

We provide example results on all datasets using our SISO and MISO architectures with different loss functions in Figure 2[2]. Each row from top to bottom shows results for FRGC, ChokePoint and YouTube Faces datasets respectively. As can be seen the use of multiple frames creates more visually appealing results. Also, the images super-resolved using the network trained to minimize SSIM loss along with the reconstruction loss produce in sharper looking images.

Comparative results on FRGC, ChokePoint and YouTube Faces datasets are displayed in Figures 3, 4 and 5 respectively. Note that horizontal lines on VSRnet results are due to the preprocessing stage of the algorithm that uses patches [3]. Since we computed our evaluation measures on a bounding box around the face in all our experiments, the spurious lines were not included in the evaluation.

[2]Since our network is trained on gray-scale images, color images are obtained by super-resolving the luminance channel and adding bicubic-upsampled chrominance channels to the result.

[3]The displayed images are cropped from left and right for better visualization. Vertical black lines also appear in cropped out areas.

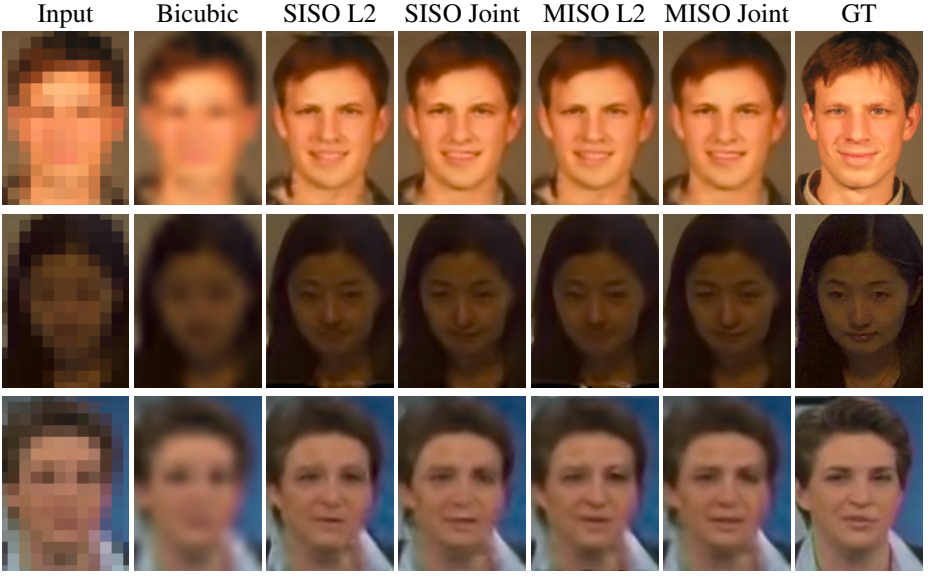

Figure 2: Qualitative results using MISO and SISO architectures with different loss functions. Each row from top to bottom contains results from FRGC, ChokePoint and YouTube Faces datasets respectively.

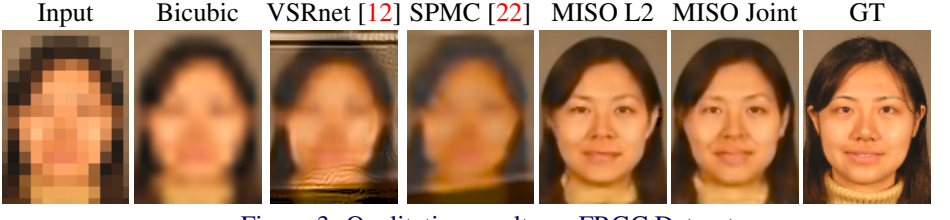

Figure 3: Qualitative results on FRGC Dataset

Figure 4: Qualitative results on ChokePoint Dataset

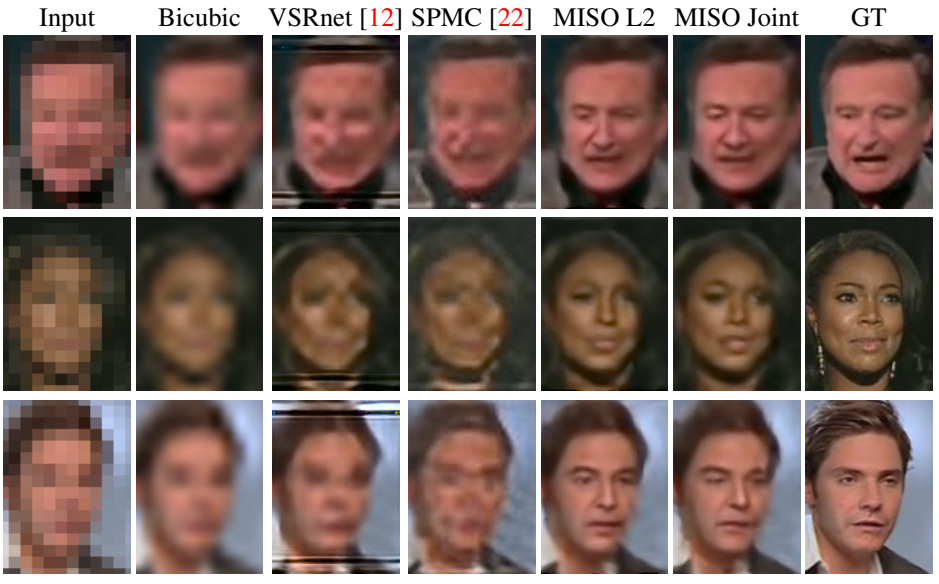

Figure 5: Qualitative results on YouTube Faces Dataset

4 Conclusion

We presented a method for super-resolving face images by using multiple images as input. Motivated by availability of cameras and saving medium, we followed a multi-input-single-output approach. The LR faces we target in this work are tiny with a facial area of around 6×6 pixels that are difficult to identify even with human eye. Our experiments show that using multi-input-single-output framework creates more accurate images compared to single-input-single-output framework.

Our network architecture consisted of two subnetworks. The first subnetwork super-resolved each frame in the sequence. This was followed by a fusion network, which used the accumulated information over the frames and produced the super-resolution of one of the frames in the sequence. We used a super-resolution network architecture that consisted of 5 layers. However, our architecture is modular and the super-resolution network can be changed with any existing super-resolution network from the literature. Thus, our fusion network idea is a general one that can be utilized with any other existing SR technique.

In our experiments, we used a sequence size of 8, which was decided empirically. The results degraded as we increased the sequence size. This might be due to the fact that our network architecture cannot handle large motion variation since we kept kernel sizes small in order to perform SR and fusion in small patches. Thus, magnification factor, frame rate and motion variation of the videos are factors that would affect the choice of sequence size. Although, increasing sequence size would provide more information for SR of a single image, it requires a more careful and sophisticated use of the data.

An important strength of the method is its independence from face models and motion estimation. Estimating motion or detecting facial landmarks become more difficult as the target face size gets smaller. An interesting extension to the current algorithm can be performing SR in a cascaded way and carry out motion estimation or model fitting when the facial details are visible in a cascade level. However, training a cascaded system (e.g. $2\times$

upsampling in each cascade) would be more time consuming and accuracy of the results in each cascade would be more critical as later stages depend on it.

This work assumes that faces were tracked throughout the sequence. In real life cases, the cameras are mounted at a specific location (e.g. surveillance cameras at an airport) and other body parts of the person are also visible in the scene. Therefore, tracking faces can benefit from human tracking. Moreover, in some surveillance settings there are certain trajectories that people follow that will also help improve tracking performance as a motion prior. Repeated motion patterns in the sequences will be also useful for our fusion network to provide a more precise output. Solving tracking and super-resolution problems in alternating steps will be the future extension of this study.

Acknowledgements

We thank Alan Sullivan, Ziming Zhang, Suzuki Daisuke and Toyoda Yoshitaka for insightful discussions.

References

[1] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1167–1183, 2002.

[2] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1167–1183, 2002.

[3] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. Attention-aware face hallucination via deep reinforcement learning. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] David Capel and Andrew Zisserman. Super-resolution from multiple views using learnt image models. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2001.

[5] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344, 2004.

[6] Sina Farsiu, Michael Elad, and Peyman Milanfar. Multiframe demosaicing and super-resolution of color images. *IEEE Transactions on Image Processing*, 15(1):141–159, 2006.

[7] G Hinton, N Srivastava, and K Swersky. RMSProp: Divide the gradient by a running average of its recent magnitude. *Neural networks for machine learning, Coursera lecture 6e*, 2012.

[8] Patrik Huber, William Christmas, Adrian Hilton, Josef Kittler, and Matthias Rätsch. Real-time 3D face super-resolution from monocular in-the-wild videos. In *Proc. SIGGRAPH*, page 67. ACM, 2016.

[9] Patrik Huber, Philipp Kopp, William Christmas, Matthias Rätsch, and Josef Kittler. Real-time 3D face fitting and texture fusion on in-the-wild videos. *IEEE Signal Processing Letters*, 24(4):437–441, 2017.

[10] Michal Irani and Shmuel Peleg. Image sequence enhancement using multiple motions analysis. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 216–221. IEEE, 1992.

[11] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016.

[13] Yang Li and Xueyin Lin. Face hallucination with pose variation. In *Inter. Conf. on Automatic Face and Gesture Recognition*, 2004.

[14] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, pages 531–539, 2015.

[15] Ce Liu, Harry Y. Shum, and Changshui Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2001.

[16] Ce Liu, Heung-Yeung Shum, and William T. Freeman. Face hallucination: Theory and practice. *Int'l J. Computer Vision*, 75(1):115–134, 2007.

[17] Ziyang Ma, Renjie Liao, Xin Tao, Li Xu, Jiaya Jia, and Enhua Wu. Handling motion blur in multi-frame super-resolution. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 5224–5232. IEEE, 2015.

[18] Kamal Nasrollahi and Thomas B Moeslund. Super-resolution: a comprehensive survey. *Machine vision and applications*, 25(6):1423–1468, 2014.

[19] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 947–954, 2005.

[20] Lyndsey C Pickup, David P Capel, Stephen J Roberts, and Andrew Zisserman. Bayesian image super-resolution, continued. In *Proc. Neural Information Processing Systems (NIPS)*, pages 1089–1096, 2007.

[21] Matan Protter and Michael Elad. Super resolution with probabilistic motion estimation. *IEEE Transactions on Image Processing*, 18(8):1899–1904, 2009.

[22] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, 2017.

[23] Oncel Tuzel, Tim K Marks, and Salil Tambe. Robust face alignment using a mixture of invariant experts. In *Proc. European Conf. Computer Vision (ECCV)*, pages 825–841, 2016.

[24] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. A comprehensive survey to face hallucination. *Int'l J. Computer Vision*, 106(1):9–30, 2014.

[25] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 529–534, 2011.

[26] Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, and Brian C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 81–88. IEEE, June 2011.

[27] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *Proc. European Conf. Computer Vision (ECCV)*, pages 372–386, 2014.

[28] Tomonari Yoshida, Tomokazu Takahashi, Daisuke Deguchi, Ichiro Ide, and Hiroshi Murase. Robust face super-resolution using free-form deformations for low-quality surveillance video. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 368–373. IEEE, 2012.

[29] Jiangang Yu and Bir Bhanu. Super-resolution of facial images in video with expression changes. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 184–191. IEEE, 2008.

[30] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *Proc. European Conf. Computer Vision (ECCV)*, pages 318–333, 2016.

[31] Xin Yu and Fatih Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 3760–3768, 2017.

[32] Qiangqiang Yuan, Liangpei Zhang, Huanfeng Shen, and Pingxiang Li. Adaptive multiple-frame image super-resolution based on u-curve. *IEEE Transactions on Image Processing*, 19(12):3157–3170, 2010.

[33] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017.

[34] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Learning face hallucination in the wild. In *Proc. of the AAAI Conf. on Artificial Intelligence*, 2015.

[35] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *European Conference on Computer Vision*, pages 614–630. Springer, 2016.