一种监控视频人脸图像超分辨技术

王嫣然[1,2], 罗宇豪[1,2], 尹 东[1,2]

[1]中国科学技术大学信息科学技术学院, 安徽 合肥 230027;
[2]中国科学院电磁空间信息重点实验室, 安徽 合肥 230027

摘要 由于目前监控视频所拍摄的人脸图像目标较小、难以辨识,图像超分辨处理已成为亟待解决监控视频图像实际应用问题的技术和手段。提出了一种针对室外监控视频人脸图像的超分辨技术,利用先验知识设置图像训练集,并进行图像空间转化、去噪等预处理操作;设计八层卷积神经网络并对各层类型及连接方式进行设定,同时设定激活函数类型及各层间传递方式函数;初始化参数并根据训练集训练网络;根据损失函数反向调整卷积核和偏置参数,完成图像输出。经过大量实际监控视频图像测试,并将本文方法和现有其他方法做对比,实验结果表明,本文方法在图像超分辨效果和处理速度上均有一定的优势。

关键词 图像处理; 图像超分辨; 卷积神经网络; 监控视频
中图分类号 TN911.73 **文献标识码** A
doi: 10.3788/AOS201737.0318012

A Super Resolution Technology of Face Image for Surveillance Video

Wang Yanran[1,2], Luo Yuhao[1,2], Yin Dong[1,2]

[1] *School of Information Science Technology, University of Science and Technology of China, Hefei, Anhui* 230027, *China*;
[2] *Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei, Anhui* 230027, *China*

Abstract Face targets in image taken by the current surveillance video are small and difficult to identify. The image super resolution processing has become the technology and means to solve the image practical application problems of surveillance video. A super resolution technology for outdoor surveillance video face image is proposed. The prior knowledge is used to construct the image training set, and some pre-processing operations like the image space conversion and denoising are operated. The convolutional neural network with eight layers is designed, and its layer types and connection mode are set. Meanwhile, the activation function types and the transmission mode functions among layers are set. The network parameters are initialized and the network is trained according to the training set. The convolution kernels and bias parameters are adjusted reversely by the loss function, and the image output is implemented. Through a large number of actual monitoring video image tests, and compared with other existing methods, the experimental results show that the proposed method has certain advantages in the effect of image super resolution and processing speed.

Key words image processing; image super resolution; convolutional neural network; surveillance video
OCIS codes 100.3010; 100.4996; 100.6640

1 引 言

近年来,利用图像采集、传输、控制、显示等设备和控制软件,我国大力开展天网工程的建设工作,对城市中固定区域进行实时监控和信息记录,为强化城市综合管理、预防打击犯罪和突发性治安灾害事故提供丰富的影像资料,图像处理技术在天网工程中将发挥巨大的应用作用。但目前我国大部分地区使用的摄像设备所拍摄的图像目标均较小,给用户使用造成极大不便,因此,对这些视频中的人脸、号牌等重要关注目标进行

收稿日期: 2016-09-02; **收到修改稿日期**: 2016-12-22
基金项目: 安徽省科技厅项目(1401b042001)
作者简介: 王嫣然(1993—),女,硕士研究生,主要从事图像处理方面的研究。E-mail: yanmmran@mail.ustc.edu.cn
导师简介: 尹 东(1965—),男,硕士,副教授,主要从事图像处理方面的研究。
E-mail: yindong@ustc.edu.cn(通信联系人)

超分辨处理，具有重要的现实意义。

光学超分辨技术突破了光学衍射极限，将显微镜的分辨率从几百纳米提高到几十纳米。目前主流的超分辨荧光显微成像技术主要分为两大类：1）基于点扩散函数调制的超分辨显微成像方法——受激辐射损耗（STED）荧光显微技术[1]和结构光照明荧光显微（SIM）技术[2]；2）基于单分子定位技术的超分辨显微成像方法——光激活定位荧光显微（PALM）技术[3]和随机光学重构荧光显微（STORM）技术[4]。STED技术利用了荧光饱和与激发态荧光受激损耗的非线性关系。SIM技术利用调制光源照明样品，将原本不可分辨的高空间频率信息编码入荧光图像中，结合计算解码获取高分辨率信息，从而提高分辨率。PALM技术和STORM技术采用牺牲时间换取空间频域的策略，并利用单分子成像加上图像重构算法实现分辨率的提升。光学成像技术的目标是突破光的物理衍射极限，其观察对象的结构是未知的，没有先验信息也无法进行训练。

与光学超分辨技术不同，图像超分辨（SR）技术[5]是提高图像分辨率的软件算法，是指对输入的单幅或多幅低分辨率（LR）图像重建高分辨率（HR）图像，从而使HR图像包含LR图像中不存在的高频细节。SR技术可以对图像进行放大处理，增大图像尺寸，并且具有去模糊、去噪和恢复图像高频细节等特征。近年来，SR技术发展迅速，诸多学者在单幅图像超分辨（SISR）技术方面都做了大量研究。然而单幅图像的简单特征不足以完整地构建出HR图像，随着机器学习在图像领域中的应用越来越活跃，使用大量数据进行训练的SISR技术开始进入研究者们的视野，并对基于外部样例学习的SISR技术进行了研究。Peleg等[6]将预测模型和稀疏表示方法结合，实现了SISR算法，取得了不错的效果，但对图像的高频细节部分并没有实现很好地修复；Kim等[7]将自然图像大梯度的稀疏性应用到SISR中，降低了计算量同时实现了对LR图像的归一化处理；Yang等[8]对上述方法进行了全面的调研和评估，认为基于样例的方法是目前最有效的SISR方法。

本文就目前监控视频图像中的人脸，以视频监控图像为训练集，经过多次迭代，训练出一个面向监控视频人脸的卷积神经网络（CNN）。经过大量实际数据的测试发现，该神经网络对于监控视频人脸超分辨具有更好的效果和普适性，将为公安机关或其他用户提供强有力的实用工具。

2　基本原理

2.1　图像超分辨技术

SISR技术的目标是从一幅LR图像中重建一幅HR图像。该问题是一个不适定问题，具有无穷多个解，因此需要一个强有力的先验知识约束解空间来得到唯一解。

根据图像的先验知识，SISR算法可以分为：预测模型方法、基于边缘方法、图像统计方法和基于样例方法。预测模型方法是通过将LR图像输入到预先定义假设的数学公式中，再输出生成HR图像，该方法没有训练数据，如双三次插值[9]、Lanczos插值[10]等。基于边缘方法是利用图像的边缘信息对图像进行超分辨重建，得到重建后的HR图像具有适当锐度的高质量边缘和更少的人为痕迹。很多SISR算法通过对边缘特征，如深度和宽度、梯度分布参数[11-12]等进行学习来重建HR图像。图像统计方法是利用图像的各种属性来从LR图像中预测出HR图像。基于样例方法是从大量的LR和HR图像对中学习，得到LR和HR图像间的非线性映射关系函数，并利用学习到的函数处理输入的LR图像得到HR图像。

2.2　卷积神经网络

CNN是一个多层的神经网络，每层由多个二维平面组成，每个平面由多个独立神经元组成。图1中f_x表示一个可训练的滤波器，b_x表示偏置，W_x+1表示权值，C_x表示卷积层，S_x表示采样层。

卷积神经元每个隐层的单元提取上一层图像的局部特征，将其映射成一个平面，特征映射函数采用Sigmoid函数作为卷积网络的激活函数，使得特征映射具有位移不变性。每个神经元与前一层的局部感受野相连。同一平面层的神经元权值共享，有相同程度的位移、旋转不变性。池化层位于卷积层之后，对输入进行降采样，常用的池化是对每个滤波器的输出求平均值或最大值。CNN通过局部感受野、共享权值和降采样的方式保证图像对位移、缩放、扭曲的稳健性。

简言之，CNN是由特征提取层（C层）和特征映射层（S层）之间交替构成的网络。C层每个神经元的输

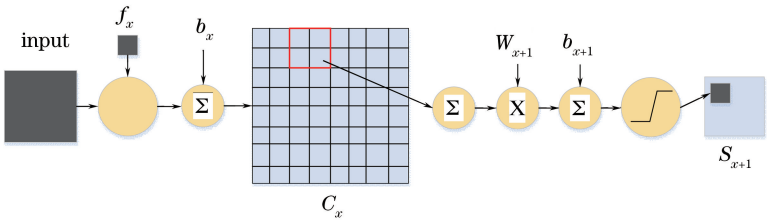

图 1　传统 CNN 结构图

Fig. 1　Traditional CNN structure

入与前一层的局部感受野相连，并提取该局部的特征，一旦该局部特征被提取后，它与其他特征间的位置关系也随之确定下来；S 层网络的每个计算层由多个特征映射组成，每个特征映射为一个平面，平面上所有神经元的权值相等，同时选用不同的函数作为激活函数用于不同的网络类型。由于一个映射面上的神经元共享权值，因而减少了网络自由参数的个数，降低了网络参数选择的复杂度。

3　本文方法

本文提出并设计、训练生成了一种面向监控视频人脸超分辨的深度 CNN，它包括设定训练集和图像预处理、CNN 参数设定、迭代训练三大部分，其工作流程如图 2 所示。

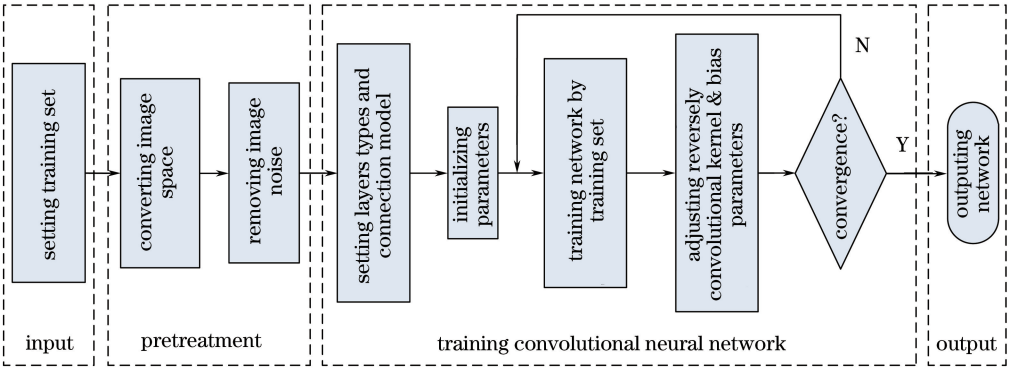

图 2　监控视频人脸超分辨深度 CNN 图

Fig. 2　Deep CNN structure for face super resolution of surveillance video

3.1　训练集设定和预处理

基于学习的 SISR 技术，其先验知识为训练样本集中的 LR 与 HR 图像对。HR 图像生成 LR 图像：

$$L = DBH + N, \tag{1}$$

式中 L 表示 LR 图像，H 表示 HR 图像，D 表示降采样算子，B 表示模糊算子，N 表示加性噪声。

共选取了来自真实监控视频的 6000 张图片作为训练集。考虑到摄像头高低因素，训练集共选用了 6 种不同大小的图片，每种图片各取 1000 张用于训练。同时选取原图作为 HR 图像，利用(1)式进行下采样 1/2 的图片作为 LR 图片，以 LR 图片作为输入、HR 图片作为输出生成一组数据，共 6000 组数据。

为了优化处理结果，加快 CNN 的训练速度，需要对输入图像进行预处理。将输入图像由 RGB 空间转换至 YUV 空间实现亮度与色度分离；对训练集中受到光照、噪声等干扰较大的图片进行去噪等处理，从而达到有效改善输出结果和训练速度的目的。

3.2　卷积神经网络设计

经过大量的实验，设计了八层结构的 CNN，分别为一层输入层、三层卷积层、三层采样层、一层输出层，既能够保证特征提取的有效性，又不至于产生过拟合现象。其中，输入层输入设置为 3，分别为 Y、U、V 通道，如图 3 所示。

图 3 中采用修正线性单元(ReLU)作为激活函数。该激活函数相比于 Sigmoid 函数能够更快地达到相同误差训练率。

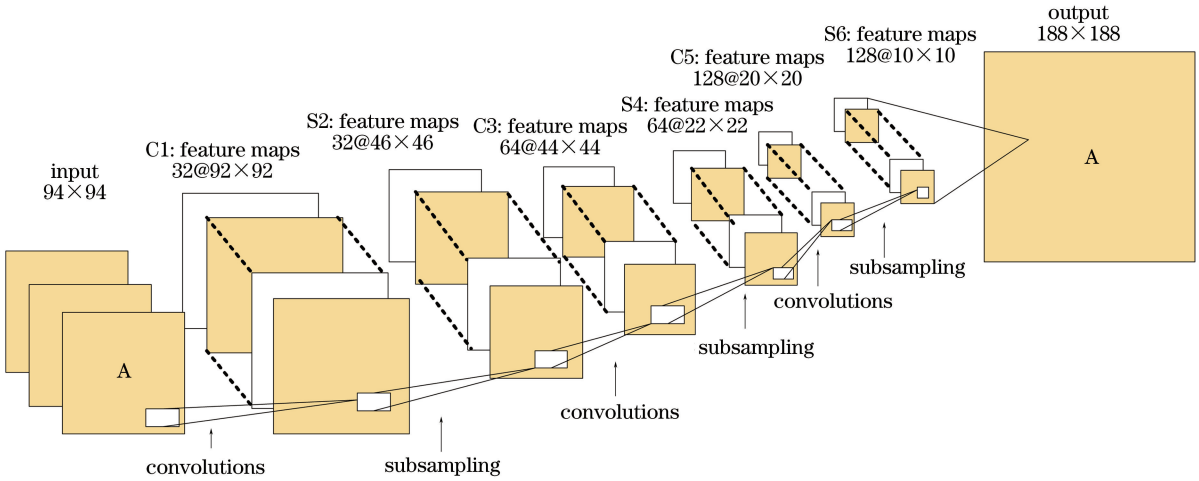

图 3　本文的 CNN 结构图

Fig. 3　Proposed CNN structure

$$F(Y)=\max(0,W*Y+B), \tag{2}$$

式中 $F(Y)$为输出，W 为卷积核，Y 为输入，B 为偏置，$*$ 表示卷积操作。

在卷积层中上一层的特征图使用可学习的卷积核进行卷积，并将卷积结果通过激活函数得到输出特征图。每个输出特征图由多个输入特征图进行组合：

$$x_j^l=f\left(\sum_{i\in M_j}x_i^{l-1}\right)*w_{ij}^l+b_j^l, \tag{3}$$

式中 x_j^l 表示第 l 层第 j 个特征图，$f(\cdot)$表示激活函数，M_j 表示输入图集合，w 表示卷积核，b 表示偏置。

采样层是输入图像采样的结果，N 个输入特征图产生 N 个输出特征图，采样过程为

$$x_j^l=f\left[k_j^l\mathrm{down}(x_j^{l-1})\right]+b_j^l, \tag{4}$$

式中 down(•)表示采样操作，k 表示可乘偏置。

第一层卷积层 C1 设置 32 个滤波器，卷积核为 3×3，偏置个数为 32，第一层卷积层为全连接，输入个数为 3、输出个数为 32，即共生成 32 个特征映射图，其中输入的 3 个神经元都采用相同权值的 32 个滤波器，称之为参数共享，能够大大降低神经网络的复杂度并加快训练速度。之后连接输入个数为 32、输出个数也为 32 的采样层 S2，每个单元与 C1 中相对应特征映射图的 2×2 邻域相连接，采样层可以防止网络出现过拟合，通过计算该特征的平均值来降低维度，即平均池化。每个特征提取后紧跟一个用来求局部平均与二次提取的采样层，这种特有的两次特征提取结构使得网络对输入样本有较高的畸变容忍能力。在采样层 S2 后再与输入个数为 32、输出个数为 64 的卷积层 C3 连接，卷积层 C3 设置 64 个滤波器，卷积核为 3×3、偏置个数为 64，卷积层 C3 和采样层 S2 的连接采用局部连接方式。将 S2 层每四个相邻的映射图作为一组输入，通过共享同一个滤波器连接至 C3 层，步长选择为 4。第一轮输入完成，将起始位置下移一位，选择每四个相邻的映射图作为输入，步长为 4。依次输入，共移位 8 次，在 C3 层生成 64 个特征映射图。在 C3 层后连接输入个数为 64、输出个数也为 64 的采样层 S4，同样采用平均池化的方式来降低维度。在采样层 S4 之后再与输入个数为 64、输出个数为 128 的卷积层 C4 连接，卷积层 C4 设置 128 个滤波器，卷积核为 3×3、偏置个数为 128，C5 层与 S4 层的连接采用局部连接方式，前 64 个特征映射图采用每两个相邻的映射图作为输入，共享一个滤波器连接至 C5 层，步长选择为 2，一轮结束后再下移一位，采用相同的方式连接至 C5 层。后 64 个特征映射图采用间隔为 1 的两个映射图作为输入，步长为 1，一轮结束后再下移两位，采用相同的方式连接至 C5 层。C5 层后连接输入个数为 128、输出个数也为 128 的采样层 S6。最后 S6 层连接输入个数为 128，输出个数为 1 的输出层，滤波器个数为 128、卷积核为 3×3、偏置个数为 128，依次连接输出，生成最后的结果。

3.3　迭代训练

根据文献[13]，利用训练集进行前向传播后，需要对初始化的各类参数进行修正，通过定义损失函数的方式反向调整各类参数，损失函数表示为

$$L(\Theta)=\frac{1}{n}\sum_{i=1}^{n}\|F(Y_i;\Theta)-X_i\|^2, \tag{5}$$

式中 $L(\Theta)$ 为损失函数，n 为训练集图像对个数，$F(Y_i;\Theta)$ 为经过网络生成的图片，X_i 为 HR 图像。

反向传播调整的过程利用随机梯度下降法的方式来最小化损失函数，迭代过程为

$$\Delta_{i+1}=0.9\cdot\Delta_i+\eta\cdot\frac{\partial L}{\partial w_i^l}, \tag{6}$$

$$w_{i+1}^l=w_i^l+\Delta_{i+1}, \tag{7}$$

式中 i 为迭代次数，l 为网络第 l 层，L 为损失函数，η 为学习率，w 为卷积核，Δ 为中间变量差值。为促进网络的收敛，η 最后一层设置为 10^{-5}，其余层为 10^{-4}。

通过将训练集的所有图像对输入并多次迭代后，整个 CNN 趋向收敛，形成最终网络。

4　实验结果与分析

实验平台为 CPU：i7，显卡：GTX960，内存：8G，操作系统：Windows 7，编程软件：Visual Studio 2013。测试集为大量监控视频中不同环境、角度和尺寸的 3000 幅图片，图 4 展示了其中的 10 幅。

图 4　原始监控视频人脸图像

Fig. 4　Original surveillance video face image

将图像放大四倍作为实验结果，并与双三次插值、超分辨率卷积神经网络（SRCNN）[13]、基于样例[14]以及 Waifu2x 方法得到的结果进行对比，来验证本文算法的有效性。双三次插值算法是传统的插值放大算法，学者们常用该算法作为基准比对算法；基于样例方法是经典的基于学习的 SISR 方法；SRCNN 和 Waifu2x 方法是和本文方法相似的利用 CNN 实现图像的超分辨算法。对比结果如图 5 所示，可以看出本文算法视觉效果最好。

针对监控视频采集到的人脸图像进行超分辨处理，无高分辨率的图像作为对比，常用的有参考评价指标如峰值性噪比（PSNR）、结构相似性（SSIM）等不适合作为本文结果的评价指标。为此，选用一种无参考的图像评价标准 JPEG2000[15]用以对各类算法超分辨后的图像进行评价，其值分布于 0～100，数值越大代表质量越好。针对各类算法生成的图像见图 5，利用 JPEG2000 得到的放大四倍后图片的评价参数如表 1 所示。

表 1　各类算法 JPEG2000 参数对比

Table 1　Comparison of JPEG2000 parameters for each algorithm

Face image	Bicubic	Example	SRCNN	Waifu2x	Proposed algorithm
1	79.9970	79.9931	79.9841	79.9900	**79.9988**
2	79.9832	79.9744	79.9339	79.9234	**79.9886**
3	80.0114	80.0062	80.0071	80.0094	**80.0186**
4	80.0167	**80.0174**	80.0150	80.0145	80.0159
5	79.9898	79.9944	79.9623	79.9836	**79.9983**
6	76.9502	74.1129	73.3986	76.4161	**79.4052**
7	79.6171	79.3733	78.5411	79.5731	**79.7684**
8	79.9618	79.9217	79.7342	79.9147	**80.0094**
9	68.9910	64.6148	65.0842	67.3550	**75.5165**
10	76.8448	75.3059	75.4171	76.6609	**77.0190**

为进一步表明本文算法的有效性，刻画了各类算法的放大两倍和四倍的 JPEG2000 参数均值统计，分别如图 6 及表 2 所示，可以看出本文算法效果最好。

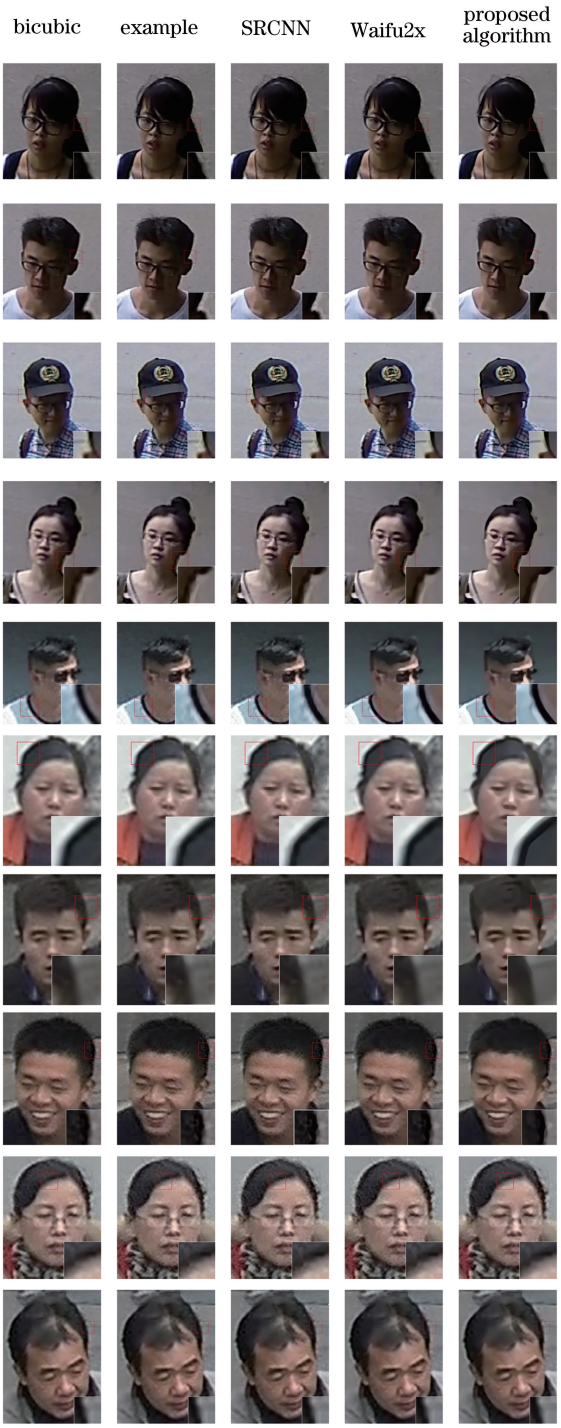

图 5　放大四倍结果

Fig. 5　Results of four times magnified

表 2　各类算法不同放大倍数时的参数均值

Table 2　Parameter mean values of different magnification times for each algorithm

Amplification rate	Bicubic	Example	SRCNN	Waifu2x	Proposed algorithm
2	58.93302	61.74297	59.74860	59.73581	**63.84919**
4	78.09354	77.65543	77.54382	78.07678	**78.74216**

再者，考虑到超分辨技术在实际中的应用，各类算法实际的运行速度也是评价算法优劣的指标。对各类算法在同一平台、同样输入以及同样放大四倍的情况下对运行速度的均值进行评估，结果如表 3 所示。

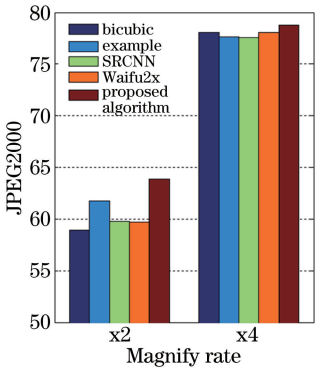

图 6　各类算法不同放大倍数时的参数均值

Fig. 6　Parameter mean values of different magnification times for each algorithm

表 3　各类算法的运行时间

Table 3　Running time for each algorithm　　s

Bicubic	Example	SRCNN	Waifu2x	Proposed algorithm
0.10078	6.263527	17.22406	7.178333	5.414944

根据表 3 可知，本文算法在计算速度上低于 Bicubic 算法，而高于其他对比算法，具有一定的优势。

5　结　　论

超分辨技术是当前图像处理领域新的研究点，具有广阔的应用前景。通过对超分辨技术及 CNN 的深入研究，针对目前我国大部分地区所使用的监控设备所获取图像目标过小、分辨率不高等问题，设计并构建了一种较实用的面向监控视频人脸超分辨的深度 CNN，在一定程度上满足了公安人员或其他用户的迫切需求。大量的实验表明，所训练的深度 CNN 相比于现有的算法，在监控人脸图像超分辨方面具有更好的效果和较强的普适性。诚然，人脸图像的进一步有效放大和放大后的去模糊化将是本文进一步的工作。

致谢　感谢合肥市公安局网络安全保卫支队提供的设备和数据支持。

参　考　文　献

[1] Hell S W, Wichmann J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy[J]. Optics Letters, 1994, 19(11): 780-782.

[2] Gustafsson M G L. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy [J]. Journal of Microscopy, 2000, 198(2): 82-87.

[3] Betzig E, Patterson G H, Sougrat R, *et al*. Imaging intracellular fluorescent proteins at nanometer resolution[J]. Science, 2006, 313(5793): 1642-1645.

[4] Rust M J, Bates M, Zhuang X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM) [J]. Nature Methods, 2006, 3(10): 793-795.

[5] Park S C, Park M K, Kang M G. Super-resolution image reconstruction: a technical overview[J]. IEEE Signal Processing Magazine, 2003, 20(4): 21-36.

[6] Peleg T, Elad M. A statistical prediction model based on sparse representations for single image super-resolution[J]. IEEE Transactions on Image Processing, 2014, 23(6): 2569-2582.

[7] Kim K I, Kwon Y. Single-image super-resolution using sparse regression and natural image prior [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(6): 1127-1133.

[8] Yang C Y, Ma C, Yang M H. Single-image super-resolution: a benchmark[C]. European Conference on Computer Vision. Springer International Publishing, 2014, 8692: 372-386.

[9] Keys R. Cubic convolution interpolation for digital image processing[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 1981, 29(6): 1153-1160.

[10] Duchon C E. Lanczos filtering in one and two dimensions[J]. Journal of Applied Meteorology, 1979, 18(8): 1016-1022.

[11] Fattal R. Image upsampling via imposed edge statistics[J]. Acm Transactions on Graphics, 2007, 26(3): 95.

[12] Sun J, Xu Z, Shum H Y. Image super-resolution using gradient profile prior[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-8.

[13] Dong C, Loy C C, He K, *et al*. Image super-resolution using deep convolutional networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 38(2): 295-307.

[14] Kim K L, Kwon Y. Example-based learning for single-image super-resolution [C]. Joint Pattern Recognition Symposium. Springer Berlin Heidelberg, 2008, 5096: 456-465.

[15] Sheikh H R, Bovik A C, Cormack L. No-reference quality assessment using natural scene statistics: JPEG2000[J]. IEEE Transactions on Image Processing, 2005, 14(11): 1918-1927.