基于卷积网络的帧率提升算法研究*

侯敬轩[a,b]，赵　耀[a,b]，林春雨[a,b]，刘美琴[a,b]，白慧慧[a,b]

（北京交通大学 a. 信息科学研究所；b. 现代信息科学与网络技术北京市重点实验室，北京 100044）

摘　要： 基于运动补偿的帧率提升算法是目前主要的帧率提升方法。为减小内插帧中的块效应、孔洞和遮挡问题，提高插值帧质量，提出一种基于卷积神经网络（convolutional neural network）的自学习帧率提升（frame rate up-conversion）方法。卷积神经网络用于利用两相邻帧预测待插值帧。在卷积神经网络的训练阶段，假设高帧率视频是存在的，网络参数由高帧率视频与低帧率视频训练而来。最后视频数据以低帧率视频加网络参数的形式传输，在接收端就可以利用卷积神经网络重建高帧率视频。实质上，这样做是通过增加视频发布者的负担以提供给视频接收者更多便利。对于视频点播网站来说，这是提升用户体验的重要因素。实验表明，该方案相对于传统的基于运动补偿的帧率提升算法，平均 PSNR 提升至少 0.6 dB，取得较大程度的提升，并且该方法是基于全局的帧预测方法，可以有效避免块效应、孔洞和遮挡问题。

关键词： 卷积神经网络；帧率提升；自学习

中图分类号： TP183　　**文献标志码：** A　　**文章编号：** 1001-3695(2018)02-0611-04

doi:10.3969/j.issn.1001-3695.2018.02.062

CNN-based frame rate up-conversion algorithm

Hou Jingxuan[a,b], Zhao Yao[a,b], Lin Chunyu[a,b], Liu Meiqin[a,b], Bai Huihui[a,b]

(*a. Institute of Information Science, b. Beijing Key Laboratory of Advanced Information Science & Network Technology, Beijing Jiaotong University, Beijing* 100044, *China*)

Abstract: Motion compensated-based frame rate up conversion (MC-FRUC) is a primary method for frame rate up conversion. This paper proposed a new self-learning-based frame rate up-conversion (FRUC) algorithm via CNN to decrease the block artifact, occlusion and holes problem in the interpolated frame. The role of CNN was to predict the intermediate frame by two adjacent frames. This article assumes that the high frame rate sequence was available in the training phase. It trained the network parameters using high frame rate and low frame rate video. Finally, it stored or transferred the data in the form of video plus network parameters. In fact, this made the video provider bear larger burden to achieve the convenience of the video receivers. This was the key to improve the user experience for the video site. In experiments using benchmark image sequences, the proposed algorithm improves the average peak signal-to-noise ratio of interpolated frames at least 0.6 dB when compared to conventional motion estimation algorithms. And the proposed method can effectively avoid the block effect, hole and occlusion problems thanks to this approach is global prediction-based method.

Key words: CNN; FRUC; self-learning-based

0　引言

随着科学技术的突飞猛进，人们对视频的质量要求不断提高，视频技术不断朝着高分辨率、高帧率的方向发展，这使得视频数据量增长极其迅猛。根据诺基亚贝尔实验室下属的咨询部门发布的一份研究报告显示，到 2020 年，音视频数据流量在数据流量增量中占比将达到 79%。激增的视频数据的存储和传输，对基础设施带来巨大挑战。因此对源视频进行压缩，包括帧率和每帧图像的压缩，成了折中的方案。

压缩后视频的质量有所降低，所以对压缩后视频进行后处理、提高视频质量就很有必要。由此实现网络带宽和海量数据存储的限制与人们观影需求之间的平衡。视频的后处理包括帧率提升（frame rate up-conversion，FRUC）算法和每一帧图像的后处理算法。本文主要讨论帧率提升算法。FRUC 技术的本质是基于前后帧的插值过程，现有的方法基本可分成两类。一类是基于相邻、连续的视频帧之间相关像素的线性插值，包括帧重复与帧平均算法[1]。这类算法的优点是算法复杂度低，缺点是由于视频中物体的运动，插值帧的效果较差。另一类算法是基于运动补偿的 FRUC（MC-FRUC）算法[2-8]。这类算法考虑物体运动，其原理如图 1 所示，所获得的插值帧的质量取决于运动估计的精度。算法包括两步：运动估计（ME）和运动补偿插值（MCI）。该类算法的优点是插值帧质量相对于第一类算法较高，缺点是计算复杂度较高。随着计算机计算能力的提升，这类算法在实际应用中逐步成为主流。

收稿日期： 2016-09-21；**修回日期：** 2017-01-03　**基金项目：** 国家自然科学基金资助项目（61402034，61210006，61501379）；北京市自然科学基金资助项目（4154082）；中央高校基本科研基金资助项目（2015JBM032）；国家科技攻关计划资助项目（2016YFB0800404）

作者简介： 侯敬轩，男，山东济南人，硕士研究生，主要研究方向为图像、视频的后处理（14120316@bjtu.edu.cn）；赵耀，男，教授，博士，主要研究方向为图像压缩、数字水印、跨媒体内容分析与理解、多媒体信息处理；林春雨，男，副教授，博士，主要研究方向为图像视频编解码、容错编码、3D 视频编码、立体视频匹配、多视点视频转换；刘美琴，女，讲师，硕士，主要研究方向为分形图像编码、3D 视频编码；白慧慧，女，教授，博士，主要研究方向为多描述视频编码、分布式视频编码.

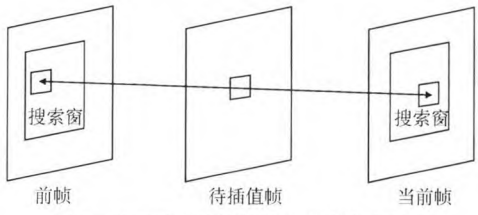


图1 基于运动补偿的帧率变换算法原理

运动估计是基于运动补偿的FRUC算法的关键步骤,基于块的运动估计(BMA)比较容易实现,所以应用最为广泛。基于块的运动估计方法得到的运动向量场,是通过比较相邻两帧相应块的残差,以残差最小的块对应的运动向量视为当前块的运动向量。在大部分时候,通过估计的运动向量并非物体真正的运动轨迹,从而产生诸如块效应、孔洞等。所以很多算法对运动向量进行后处理,通过修正运动向量得到一个相对平滑、正确的运动向量场。

基于块的运动估计主要包括单边运动估计和双边运动估计两类。用单边运动估计得到的插值帧图像存在较严重的重叠和孔洞。双边运动估计充分利用参考帧与当前帧之间的空间对称性,可以在一定程度上避免重叠和孔洞效应,但代价是运算量的增加。

本文提出一种基于卷积神经网络(convolutional neural network,CNN)的FRUC算法,由于本文算法使用全局预测,避免了基于块的运动估计,所以插值帧不存在重叠和孔洞效应。另外,在视频点播等应用场景下,接收到的视频已被压缩过,基于运动估计的FRUC算法通过非学习的方式实现,割裂了插值帧与源视频帧的关系。相反,因为本文假设源视频已知,对应的卷积网络由源视频训练而来,所以本文算法可以充分利用源视频的信息。实验结果表明,本文算法的结果优于很多传统的基于运动补偿的算法。

1 基于CNN的FRUC算法

本文算法的目的是构建一个非线性映射,把现有两相邻帧映射到待插值帧,可表示为

$$F_{n+0.5}=\varphi(F_n,F_{n+1}) \tag{1}$$

其中:φ 是非线性映射;F_n、F_{n+1} 是低帧率视频的两相相邻帧;$F_{n+0.5}$ 是插值帧。

1.1 简介

本文中,上述非线性映射 φ 用CNN实现。CNN是人工神经网络的一种,已成为当前语音分析和图像识别领域的研究热点。其使图像可以直接作为网络的输入,避免了传统识别算法中复杂的特征提取和数据重建过程。它的权值共享网络结构受生物神经网络启发,降低了网络模型的复杂度,减少了权值的数量。该优点保证了在相同参数数量情况下有更好的拟合效果。基于此,选择CNN作为映射函数,使得可以忽略网络参数在存储或传输过程中造成的数据增量。

文献[9,10]中构造了一个三层的卷积网络,实现了图像超分辨率重建,其网络结构如图2所示。它包括两个隐层,网络输入为低分辨率图像,输出为重建的高分辨率图像。每一层的处理过程可表示为

$$\begin{aligned} &F_l=LR && l=0\\ &F_l=\max(0,W_l\times F_{l-1}+B_l) && l=1,2\\ &F_l=W_l\times F_{l-1}+B_l=HR && l=3 \end{aligned} \tag{2}$$

其中:LR 表示用双三次插值上采样后的低分辨率图像,以保持输入图像和输出图像大小相近;W_l 和 B_l 分别表示各层的权值和偏置,HR表示重建的高分辨率图像。该网络的每一层可看做一个算子:

a)第一层为特征表示。用 n 个 9×9 大小的卷积对输入层进行卷积,得到一个高维向量。该高维向量由 n 个特征图组成,构成对输入图像块的高维表示。

b)第二层为非线性变换。用 m 个 1×1 大小的卷积核卷积第一层的高维向量,得到另一个高维向量。该高维向量由 m 个特征图组成,构成输入图像的另一组高维表示。

c)第三层为重建。用一个 5×5 的卷积核对第二层的特征图像进行卷积,得到最终的重建图像。该图像就是所希望得到的近似于原图像的结果。

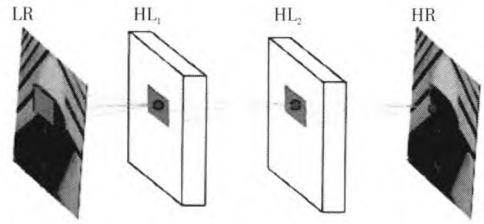


图2 用于图像分辨率重建的网络结构

可见,学习非线性映射 φ 就是估计卷积网络的参数 $\{W_1,W_2,W_3;B_1,B_2,B_3\}$。参数的估计通过最小化网络的重建结果 $\varphi(Y;\theta)$ 与原高分辨率图像 X 之间的损失函数得到。给定一个高分辨率图像集 $\{X_i\}$ 与对应的低分辨率图像集 $\{Y_i\}$,用均方误差(MSE)作为损失函数,网络的训练过程可描述为

$$\min_{\theta}\frac{1}{n}\sum_{i=1}^{N}\|\varphi(Y_i;\theta)-X_i\|^2 \tag{3}$$

其中:N 是训练样本的数量,损失函数的最小化通过随机梯度下降法实现。

需要说明的是,在训练阶段,不是直接把训练集图像作为样本,而是从训练集中随机提取图像块作为训练样本。实验表明,适当尺寸的训练样本更有利于重建效果。该算法取得了优于传统的超分辨率重建算法的效果,在边缘保持和降噪方面的效果尤为突出。

1.2 视频未压缩时的FRUC算法

借鉴上述网络,对插值帧进行估计,网络结构如图3所示。区别于参考网络的是,其以两相邻帧作为输入来预测待插值帧。网络的训练过程可表示为

$$\min_{\theta}\frac{1}{n}\sum_{n=1}^{N}\|\varphi(F_n,F_{n+1};\theta)-X_{n+0.5}\|^2 \tag{4}$$

其中:$X_{n+0.5}$ 为待插值帧对应的高帧率视频中的帧图像,为卷积网络代表的非线性映射。插值过程由式(1)表示。训练方法与网络参数与参考网络保持一致。

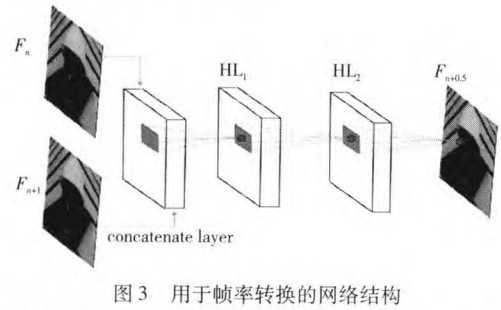


图3 用于帧率转换的网络结构

本文假设源高帧率视频是可获得的,根据每一个需要帧率

下采样后进行传输的视频训练一个网络,然后以视频加网络参数的形式传输,在接收端重建高帧率视频,以克服直接传输高帧率视频对流量或带宽带来的压力。之所以针对每一个视频训练一个网络,是因为这样做可以避免通过增加网络层数以提升重建效果的做法。网络层数增加必然带来计算量的增加,处理速度变慢,这对一些互联网终端来说是不现实的。在实际应用中,视频发布方往往存在富余的计算能力。所以这样做,实际是通过增加视频发布者的负担,以提供给视频接受者更多便利。这对于一些视频点播网站来说是有意义的,因为让用户以最小的代价提高观影质量,可以提升用户体验,这是企业经营成功的关键。

1.3　视频经过压缩时的 FRUC 算法

虽然本文算法避开使用基于块的运动补偿算法,但是因为物体运动的影响,初步得到的插值帧也存在幻影效应。尤其是视频压缩后,帧图像会产生模糊、变形等现象。利用压缩后视频进行帧率提升,插值帧的质量大大降低。本文针对压缩后视频采取了一种双网络模型。首先,采用 1.2 节中的策略,训练 CNN1 用于视频的帧率提升;其次,模仿传统的基于运动补偿算法对运动向量的后处理,本文训练卷积网络 CNN2 用于帧率提升后视频的后处理。

本文同时考虑压缩造成的质量下降与插值带来的幻影效应。使用 CNN 构造一个从插值初步得到的高帧率视频到原高帧率视频的映射。可以把该卷积网络视做滤波器,其起到增强图像质量的作用。此处,选择一个两层网络,其结构如图 4 所示。与参考网络不同的是,去除参考网络的隐层 HL2,因为卷积网络的运算时间主要集中在该层,实验表明,两层网络的处理时间是三层网络的十分之一左右,但重建效果相近。所以,从实际应用的角度出发,选择两层的卷积网络,进行预估计帧的后处理。该网络的训练过程可表示为

$$\min_{\theta}\frac{1}{n}\sum_{n=1}^{N}\|\varphi'(F_n;\theta')-X_n\|^2 \tag{5}$$

处理过程表示为

$$F'_n=\varphi'(F_n;\theta') \tag{6}$$

其中:F'_n 为处理后的帧,F_n 为插值后视频的帧,X_n 为对应的高帧率视频中的帧,φ'为卷积网络代表的非线性映射。

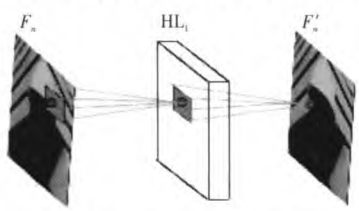


图 4　用于估计帧后处理的网络结构

综上所述,在现实场景中,视频压缩传输是普遍现象,所以本文针对视频压缩后的情况,如图 5 所示,对本文的总体思路进行概括。

2　实验结果

实验选取了多个 4∶2∶0的 CIF 格式的 YUV 标准测试序列,来验证本文算法的有效性。这些序列各有特点,可以比较全面地反映算法的优劣。这些特点包括如 mobile 的慢速移动和丰富细节、coastguard 的水平运动、foreman 的前景运动差异明显等。本文根据序列是否压缩,分别对未压缩和压缩两种情况采用不同的策略。另外,鉴于人眼对颜色分量不敏感,本文只对亮度分量(*Y* 分量)进行处理。

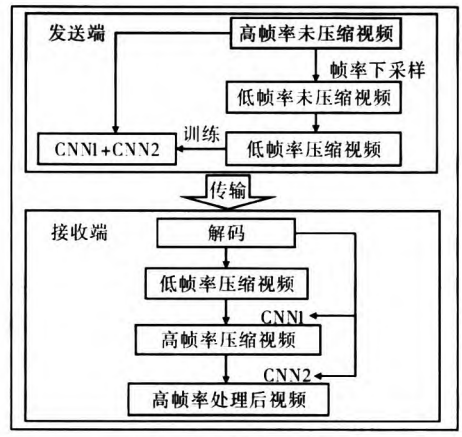


图 5　本文算法流程图

卷积网络的参数设置:CNN1 的三个卷积层的卷积核个数分别为 64、32、1,卷积核大小分别为9×9、1×1、5×5。CNN2 的两个卷积层的卷积核个数分别为 64、1,卷积核大小分别为9×9、5×5。样本对的大小为 36×36(输入)与 24×24(输出)。本文利用 Caffe 平台进行模型训练,利用 MATLAB 进行插值处理与插值后视频的处理。

本文对每一个视频训练一个网络,这无疑会有很大的工作量。在实际操作中,首先在多个视频中随机抽取 10 万对训练样本进行预训练,然后再次在此基础上针对每一个视频进行进一步微调。这是 CNN 用在分类问题时的普遍做法。这样做可以降低每次训练的时间,同时提高精度。

2.1　视频未压缩时的情况

视频未压缩的情况下,使用单个网络 CNN1 进行插值。为了验证算法的有效性,本文选择多种 MC-FRUC 算法进行比较。为保证与参考文献一致,以每个序列的前 100 帧为测试样本。首先,对该 100 帧序列进行帧率下采样,然后使用 CNN1 对下采样后序列进行插值。以 PSNR 作为客观指标对不同方法进行比较。表 1、2 和图 6 分别比较不同序列下插值处理与后处理的客观效果和主观效果。

表 1　不同序列下各种算法的平均 PSNR　/dB

序列	文献[3]	文献[4]	文献[5]	本文方法
foreman	32.40	32.61	32.79	32.84
highway	31.50	31.18	31.31	34.20
mobile	23.71	22.89	26.41	32.59
news	36.51	35.60	37.58	37.07
Stefan	25.79	23.89	27.55	26.96
平均	29.98	29.23	32.13	32.73

表 2　不同序列下各种算法的平均 PSNR　/dB

序列	文献[6]	文献[7]		本文方法
		框架 1	框架 2	
Akiyo	41.05	41.94	45.11	45.71
coastguard	26.22	28.38	30.75	33.68
flower	25.83	26.81	27.93	31.67
foreman	29.49	31.54	32.71	32.84
hall	32.16	34.99	35.46	34.01
highway	28.65	31.46	32.23	34.20
mobile	21.88	22.66	26.18	32.59
mother & daughter	37.87	39.23	41.45	39.10
news	33.24	34.59	35.46	37.07
Stefan	22.67	24.03	26.32	26.96
平均	29.90	31.56	33.36	34.78

(a)原图 (b)本文算法得到的插值帧

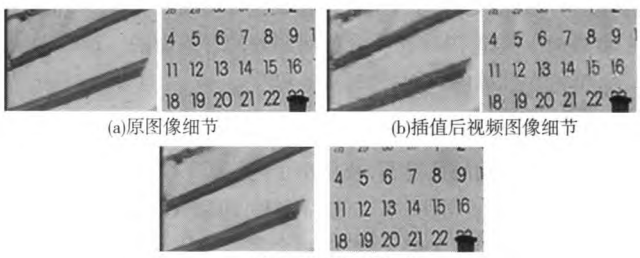
图6 本文算法插值效果

2.2 视频经过压缩时的情况

如1.2节所述,视频经过压缩的情况下,以每个序列的前100帧为测试样本,采取分步策略对视频进行处理。首先,对该100帧序列进行帧率下采样,然后对下采样后序列进行压缩得到序列 Y。压缩方法采用H.264视频编码标准,编码模式采用quality-based模式,QP设定为40。首先,运用CNN1对序列 Y 进行插值,然后利用CNN2对插值后的序列进行处理。本文以PSNR作为客观指标对插值与后处理的效果进行比较。表3和图7分别比较不同序列下插值处理与后处理的客观效果和主观效果。

表3 压缩后视频的FRUC的PSNR比较 /dB

序列	插值帧	整个序列	
		无后处理	后处理
foreman	30.39	30.98	31.44
highway	32.97	33.72	33.98
mobile	27.70	27.23	27.59
平均	30.35	30.64	31.00



(a)原图像细节 (b)插值后视频图像细节

(c)处理后的帧图像细节

图7 视频压缩时FRUC效果

3 结束语

本文研究了一种基于卷积网络的自学习FRUC算法,该算法基于图像的全局预测,可有效避免块效应与孔洞效应。另外本文算法通过自学习完成,可充分利用原视频的信息。本文根据视频是否压缩采取不同策略。在视频经过压缩的情况下,同时考虑压缩损失与插值引起的幻影效应,采取分步策略进行处理,使得视频整体质量提高。实验表明,本文算法相对多种传统算法有明显优势。需要说明的是,文中网络的训练结果受训练次数的影响较大,由于时间的限制,本文结果并非最优结果。如果增加训练时间,效果有望进一步增强。本文及参考文献只考虑了亮度分量,如果考虑色度分量,本文算法还需进一步研究。

参考文献:

[1] Chen H F, Lee S H, Kwon O J, *et al*. Smooth frame insertion method for motion-blur reduction in LCDs[C]//Proc of the 7th Workshop on Multimedia Signal Processing. 2005: 581-584.

[2] 杨爱萍, 董翠翠, 侯正信, 等. 基于多帧运动估计的帧率提升算法[J]. 计算机应用研究, 2012, 29(10): 3952-3955.

[3] Kang S J, Yoo S J, Kim Y H. Dual motion estimation for frame rate up-conversion [J]. IEEE Trans on Circuits and Systems for Video Technology, 2010, 20(12): 1909-1914.

[4] Choi B D, Han J W, Kim C S, *et al*. Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation[J]. IEEE Trans on Circuits and Systems for Video Technology, 2007, 17(4): 407-416.

[5] Wang Ci, Zhang Lei, He Yuwen, *et al*. Frame rate up-conversion using trilateral filtering[J]. IEEE Trans on Circuits and Systems for Video Technology, 2010, 20(6): 886-893.

[6] Han S C, Woods J W. Frame-rate up-conversion using transmitted motion and segmentation fields for very low bit-rate video coding [C]//Proc of International Conference on Image Processing. 1997: 747-750.

[7] Kim U S, Sunwoo M H. New frame rate up-conversion algorithms with low computational complexity [J]. IEEE Trans on Circuits and Systems for Video Technology, 2014, 24(3): 384-393.

[8] Inseo H, Jung H S, Sunwoo M H. Novel frame rate up-conversion algorithm based on prediction and recursive search[C] //Proc of IEEE Workshop on Signal Processing Systems. [S. l.]: IEEE Press, 2015: 1-4.

[9] Dong Chao, Loy C C, He Kaiming, *et al*. Learning a deep convolutional network for image super-resolution[C]//Proc of Computer Vision. 2014: 184-199.

[10] Dong Chao, Loy C C, He Kaiming, *et al*. Image superresolution using deep convolutional networks[C]//Proc of European Conference on Computer Vision. [S. l.]: Springer International Publishing, 2014:184-199.

(上接第610页)

[7] Li Jianan, Xu Tingfa, Zhang Kun. Real-time feature-based video stabilization on FPGA[J]. IEEE Trans on Circuits & Systems for Video Technology, 2017,27(4):907-919.

[8] Zhang Lei, Xu Qiankun, Huang Hua. A global approach for fast video stabilization[J]. IEEE Trans on Circuits & Systems for Video Technology, 2017,27(2):225-235.

[9] Manasa K, Channappayya S S. An optical flow-based full reference video quality assessment algorithm[J]. IEEE Trans on Image Processing, 2016, 25(6): 2480-2492.

[10] 陈滨, 杨利斌, 赵建军. 基于SIFT特征的视频稳像算法[J]. 兵工自动化, 2016,35(4): 45-48.

[11] Koh Y J, Lee C, Kim C. Video stabilization based on feature trajectory augmentation and selection and robust mesh grid warping[J]. IEEE Trans on Image Processing, 2015, 24(12): 5260-5273.

[12] 吉淑娇, 朱明, 雷艳敏, 等. 基于改进运动矢量估计法的视频稳像[J]. 光学精密工程, 2015, 23(5): 1458-1465.

[13] Zhu Juanjuan, Fan Jing, Guo Baolong. Adaptive electronic image stabilization algorithm resistant to foreground moving object[J]. Acta Photonica Sinica, 2015, 44(6): 0610002.

[14] 闫利, 陈林. 一种改进的SURF及其在遥感影像匹配中的应用[J]. 武汉大学学报: 信息科学版, 2013, 38(7): 770-773.

[15] 程德强, 郭政, 刘洁, 等. 一种基于改进光流法的电子稳像算法[J]. 煤炭学报, 2015, 40(3): 707-712.