基于三维卷积神经网络的无参考视频质量评价

张淑芳, 郭志鹏*

天津大学电气自动化与信息工程学院, 天津 300072

摘要 为了在不借助参考视频的条件下准确评价失真视频质量,提出一种应用三维卷积神经网络提取失真视频时空域特征的通用型无参考视频质量评价算法。在视频质量库上训练卷积神经网络模型 3D ConvNets,使 3D ConvNets 学习到与视频失真程度相关的特征;应用 3D ConvNets 对输入的失真视频进行特征提取,对提取得到的质量特征先后进行 L2 范数规则化和主成分分析以防止过拟合并去除冗余特征;使用线性支持向量回归根据视频质量特征预测失真视频的质量分数。实验结果表明,本文算法能够较为准确地评价多种视频失真类型,并且在更换测试视频库后依然保持较高的评价准确度,同时算法评价视频质量的计算复杂度极低。

关键词 成像系统; 视频质量评价; 无参考; 三维卷积神经网络; 时空域特征; 线性支持向量回归

中图分类号 TN919.8 **文献标识码** A **doi**: 10.3788/LOP55.071101

No-Reference Video Quality Assessment Based on Three-Dimensional Convolutional Neural Networks

Zhang Shufang, Guo Zhipeng

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract In order to assess the quality of distorted videos accurately without reference videos, a universal no-reference video quality assessment algorithm is proposed, which applies three-dimensional (3D) convolutional neural networks to extracting spatiotemporal features of distorted videos. Firstly, the convolutional neural network model 3D ConvNets is trained on the video quality database, and then the features related to video distortion degree are learned. Then, 3D ConvNets is used to extract features of the input distorted video, after which L2-normalization and principal component analysis are performed to prevent overfitting and eliminate redundancy. Finally, linear support vector regression is used to predict quality score of the distorted video based on video quality features. The experimental results show that the proposed algorithm can assess video quality accurately across different kinds of distortion, and it still maintains a high level of accuracy when the test video database is changed. Last but not least, the computational complexity of quality assessment process is extremely low for the proposed algorithm.

Key words imaging systems; video quality assessment; no-reference; three-dimensional convolutional neural networks; spatiotemporal features; linear support vector regression

OCIS codes 110.3000; 100.2000; 100.4996

1 引 言

由于视频压缩技术和视频传输信道的限制,视频中会不可避免地引入编码失真和传输失真,严重影响了视频服务用户的观看体验。主观评价方法对于测试人群、测试环境及统计方法等有非常严格的规定,实施难度较大。而客观评价方法中,无参考质量评价因为其不需要参考视频的优点,实用性较高。由于在实际系统中,视频往往包含多种失真类型[1],因此研究一种能够评价多种视频失真类型的通用型无参考视频质量评价算法逐渐成为热点。

无参考评价方法需要抓住反映视频质量的最本质特征[2]。Saad 等[3]分别在像素域和离散余弦变换(DCT)域对失真视频提取参数的统计特征,再通

收稿日期: 2017-11-27; **收到修改稿日期**: 2017-12-17

作者简介: 张淑芳(1979—),女,博士,副教授,硕士生导师,主要从事图像处理方面的研究。E-mail: shufangzhang@tju.edu.cn

* **通信联系人**。E-mail: zhipengguo@tju.edu.cn

过运动矢量计算整体运动特征和运动相干性特征，结合空间域和时间域特征评价视频质量，大大提高了仅使用图像质量评价方法评价视频质量的准确度。Li 等[4]对失真视频按视频块作三维 DCT(3D-DCT)，提取变换后参数的分布形状、子带间比例和方向等时空域特征，以评价视频的失真程度。Li 等首先对失真视频块进行三维剪切波变换[5]，提取变换后参数的初级统计特征，然后将特征输入一维卷积神经网络(CNN)中，得到性能更好的高层次特征评价视频质量[6]。王春峰等[7]将 3D 卷积神经网络引入视频质量评价中，并对模型进行修改与适配，使新网络在评价视频质量时取得了较好的效果。

上述算法虽然取得了与主观评价结果较高的一致性，但是在特征提取过程中都需要进行 DCT 或小波变换等域变换，计算复杂度较高，提取特征耗时很长。基于此，本文提出了一种基于三维卷积神经网络的通用型无参考视频质量评价算法，算法包含两部分：在模型训练阶段，通过在视频质量库上对 3D ConvNets 进行训练，完成 3D ConvNets 对质量特征的学习，建立质量评价模型；在质量评价阶段，失真视频被直接输入 3D ConvNets 中，使用 L2 范数规则化和主成分分析优化提取出的特征向量，利用线性支持向量回归(SVR)[8]建立特征向量与视频质量分数的映射关系模型，从而预测得到输入视频的质量分数。实验结果表明，本文算法对于多种视频失真类型的评价准确度均较高，对于不同测试视频库的评价性能保持稳定，同时算法质量评价过程的计算复杂度很低，评价速度快。

2 模型介绍及训练

2.1 3D ConvNets 介绍

作为深度学习模型的一种，卷积神经网络已成为语音分析和图像识别领域的研究热点。然而，二维卷积神经网络只适用于提取图像级别的特征，对于视频来说，由于二维卷积只发生于视频空间域的每一帧图像上，得到的输出也仅仅是图像级别的二维结果，因此网络在每一次卷积操作之后都会丢失视频的时域信息。基于此，Tran 等[9]提出了三维卷积神经网络模型 3D ConvNets，该模型使用的都是三维的卷积和池化算子，卷积和池化操作都发生于每一个视频块上，相较于二维卷积神经网络，3D ConvNets 更适用于提取视频的时空域特征。Tran 等通过大量实验，证明了 3D ConvNets 能够在训练过程中学习到高性能的时空域特征，从而出色地完成一系列视频处理任务。因此，本文将 3D ConvNets 应用于算法的特征提取过程中，使得输入的失真视频可以不经过域变换等复杂度高的计算就提取出高性能的质量特征。

2.2 模型训练过程

本文算法对 3D ConvNets 模型的训练过程如图 1 所示，其中 conv-表示的是卷积层，pool-表示的是池化层，fc-表示的是全连接层，softmax with loss 层是用来计算机器学习算法 softmax 根据模型学习到的特征对训练视频分类的结果与视频自带标签的差，卷积层中的数字表示卷积层中滤波器的个数，全连接层中的数字表示这一层的特征输出维数。所有卷积层的卷积算子大小都是 $2\times2\times2$，移动步长在时间域和空间域上都是 1；除了第一层池化层的池化算子大小是 $1\times2\times2$ 外，其余所有池化层的池化算子大小都是 $2\times2\times2$，所有池化层的池化算子移动步长在时间域和空间域上都是 1。

具体训练方法描述如下。

1) 根据视频质量库中训练视频的主观评价分数，将训练视频按失真程度的不同分为 10 组，并且为这 10 组视频分配 0，1，2，…，9 共 10 个不同的标签，将这些带有标签的视频输入 3D ConvNets 中。

2) 随机初始化 3D ConvNets 中卷积、池化和全连接层的各个权重及偏置参数，第一次迭代开始。输入视频在 3D ConvNets 中按视频块进行卷积和池化运算。前两个卷积层和池化层提取出的都是视频的低等级特征，比如边缘等，后面卷积、池化以及全连接层提取的是视频的高等级抽象特征。

3) 在 3D ConvNets 中经过一系列运算后，最后一层输出的 10 维特征向量及其对应的标签被输入 softmax with loss 层中。在这一层中，首先利用 softmax 根据特征向量对输入视频分类，得到预测标签，然后将预测标签与输入的标签进行比较，计算损失。

4) 第一次迭代结束，通过反向传播根据损失调整 3D ConvNets 中卷积、池化以及全连接层中的各个权重及偏置参数，之后开始第二次迭代。

5) 不断进行迭代和反向传播调整权重及偏置参数，直到输出的损失足够小为止，此时 3D ConvNets 成为已经训练好的能够有效提取失真视频质量特征的卷积神经网络模型。

对 3D ConvNets 模型的训练过程就是 3D ConvNets 自身的特征学习过程。通过这一过程，3D ConvNets 借助视频质量库学习到能够有效反映输入视频失真严重程度的特征。

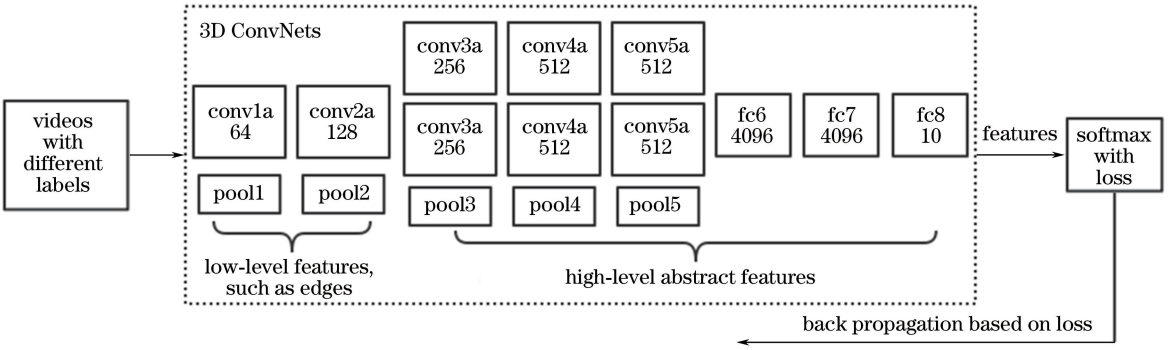

图 1 3D ConvNets 模型的训练过程

Fig. 1 Training process of 3D ConvNets model

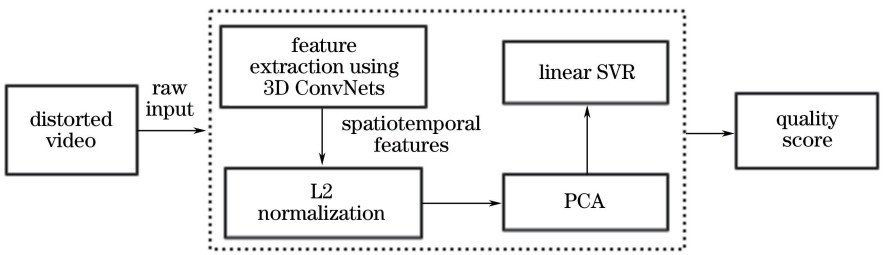

图 2 本文算法的质量评价过程

Fig. 2 Quality assessment process of the proposed algorithm

3 质量评价

算法的质量评价过程如图 2 所示，失真视频直接输入 3D ConvNets 中进行特征提取，然后对提取出的时空域特征进行 L2 范数规则化以避免出现过拟合问题，之后再对特征进行主成分分析(PCA)去除冗余特征，保留有用信息，最后通过线性 SVR 预测得到失真视频的质量分数。

具体质量评价过程如下。

1) 将每一个输入的失真视频分为帧长为 16 的视频段，每两个连续的视频段之间有 8 帧的重叠区间。将这些视频段输入 3D ConvNets 中，将 fc7 层的输出向量作为提取出来的特征。对一个失真视频所有视频段的特征取平均，得到维数为 4096 的特征向量 $\boldsymbol{A}_{\mathrm{a}}$，表示为

$$\boldsymbol{A}_{\mathrm{a}}=(x_1,x_2,\cdots,x_{4096})。\tag{1}$$

2) 对 $\boldsymbol{A}_{\mathrm{a}}$ 进行 L2 范数规则化以避免出现过拟合，计算方法为

$$\boldsymbol{F}=(y_1,y_2,\cdots,y_{4096})=\frac{(x_1,x_2,\cdots,x_{4096})}{\|\boldsymbol{A}_{\mathrm{a}}\|_2},\tag{2}$$

式中：$\|\boldsymbol{A}_{\mathrm{a}}\|_2$ 为向量 $\boldsymbol{A}_{\mathrm{a}}$ 的 L2 范数。

3) 为了去除特征向量中的冗余部分，保留有用信息，同时也是为之后与 V-BLIINDS 算法进行平等的性能比较(V-BLIINDS 算法中失真视频的特征向量维数是 46)，使用主成分分析降低 $\boldsymbol{F}$ 的维数为 46，得到最终的视频质量特征($f_1,f_2,\cdots,f_{46}$)。

4) 利用线性 SVR 预测失真视频的质量分数，实现过程分为两步：首先提取训练集中视频的质量特征，与其对应的主观评价分数一起监督训练线性 SVR 模型，建立视频特征与质量分数的映射关系；然后提取测试集视频的质量特征，将特征输入已经训练好的线性 SVR 模型中，通过映射函数计算得到最终的质量分数。

4 实验结果与性能分析

4.1 视频库及实验方法

实验采用 LIVE[10] 和 CSIQ[11] 两个视频质量数据库，其具体信息如下。

1) LIVE 视频库有 10 个参考视频，150 个失真视频。失真类型包含 MPEG-2 压缩编码失真、H.264压缩编码失真、无线传输失真和 IP 传输失真 4 种，视频分辨率为 768×432，每一个失真视频都有对应的主观评价分数。

2) CSIQ 视频库有 12 个参考视频，216 个失真视频。失真类型包含 MJPEG 压缩编码失真、H.264 压缩编码失真、HEVC 压缩编码失真、小波压缩编码失真、丢包传输失真、加性高斯白噪声传输失真 6 种，视频分辨率为 832×480，每一个失真视频都有

对应的主观评价分数。

性能评价指标采用斯皮尔曼相关系数(SROCC)和线性相关系数(LCC),SROCC 和 LCC 的值越接近 1,表明算法评价结果与主观评价结果的一致性越好,即算法的评价准确度越高。

实验首先验证算法对各单一视频失真类型的评价准确度,将视频库中每一类失真类型的所有失真视频分为视频内容互不重叠的两部分,一部分作为训练集,占 80%,另一部分作为测试集,占 20%。计算算法对测试集视频的评价分数与其主观评价分数的 SROCC 和 LCC,随机进行 7 组训练集与测试集划分方式不同的实验,取各次实验 SROCC 和 LCC 的中值作为对这一类型的评价准确度;然后用相同的方法验证算法对视频库所有视频的评价准确度。

4.2 评价准确度验证

首先在 LIVE 视频库上验证本文算法的评价准确度。为了使测试结果更加具有说服力,对全参考(FR)视频质量评价算法 PSNR、SSIM[12] 和 MOVIE[13],部分参考(RR)视频质量评价算法 VQM[14],以及无参考(NR)视频质量评价算法 V-BLIINDS[3] 进行相同的测试实验并加以比较。表 1 和表 2 分别列出了这些算法在 LIVE 视频库上测试实验得到的 SROCC 和 LCC 中值。

表 1 LIVE 视频库上测试得到的 SROCC 中值

Table 1 Median SROCC of tests on LIVE database

Type	Algorithm	Wireless	IP	H.264	MPEG	All
FR	PSNR	0.575	0.477	0.599	0.403	0.549
	SSIM	0.715	0.603	0.781	0.786	0.669
	MOVIE	0.811	0.716	0.766	0.773	0.789
RR	VQM	0.721	0.638	0.652	0.781	0.702
NR	V-BLIINDS	0.792	0.725	0.827	0.861	0.731
	Proposed	0.743	0.714	0.767	0.857	0.718

表 2 LIVE 视频库上测试得到的 LCC 中值

Table 2 Median LCC of tests on LIVE database

Type	Algorithm	Wireless	IP	H.264	MPEG	All
FR	PSNR	0.593	0.484	0.611	0.417	0.565
	SSIM	0.730	0.611	0.793	0.802	0.688
	MOVIE	0.842	0.766	0.814	0.798	0.813
RR	VQM	0.755	0.667	0.666	0.813	0.730
NR	V-BLIINDS	0.813	0.802	0.923	0.919	0.785
	Proposed	0.787	0.763	0.837	0.866	0.766

从表 1 和表 2 可以看出,本文算法对 LIVE 库中各个单一失真类型和所有失真视频的评价准确度都显著高于 PSNR、SSIM 和 VQM 算法。与当前通用型无参考视频质量评价中的主流算法 V-BLIINDS 相比,本文算法在评价准确度上也有很强的竞争力。虽然本文算法的评价准确度低于全参考视频质量评价算法 MOVIE,但是作为无参考视频质量评价算法,不需要参考视频,实用性更高是其相较于全参考视频质量评价算法的重要优势。

为了验证本文算法对不同视频库的评价性能,在 CSIQ 视频库上也进行相同的评价准确度验证实验。表 3 列出了本文算法和 V-BLIINDS 在 CSIQ 视频库上测试实验得到的 SROCC 和 LCC 中值。

从表 3 可以看出,V-BLIINDS 算法对于 CSIQ 库中失真类型为 Packet-loss 的失真视频评价准确度较低,并且当评价视频库中所有视频时准确度有明显的下降。相比之下,本文算法无论是对各单一失真类型还是所有失真视频的评价准确度都维持在较高的水平,明显超过了 V-BLIINDS 算法。分析认为出现这种情况的原因是 V-BLIINDS 算法是在 LIVE 库上比较分析各类失真视频和无失真视频统计特征差别而进行特征提取,由于 CSIQ 库和 LIVE 库的视频内容和失真类型都不相同,比如 CSIQ 库与 LIVE 库中视频拍摄的场景没有重叠,CSIQ 库包含更多拍摄快速运动场景(如打篮球和赛马)的视频,CSIQ 库比 LIVE 库多了 Packet-loss 与 Wavelet 两种失真类型等,这些不同使得 V-BLIINDS 算法中一部分基于 LIVE 库视频选定的特征对 CSIQ 库中失真视频的质量评价作用较小甚至无效。本文算法明显不受视频内容及失真多样性的影响,在改变

表 3 CSIQ 视频库上测试得到的 SROCC 和 LCC 中值

Table 3 Median SROCC and LCC of tests on CSIQ database

Distortion	SROCC		LCC	
	V-BLIINDS	Proposed	V-BLIINDS	Proposed
H.264	0.938	0.943	0.955	0.963
Packet-loss	0.426	0.829	0.518	0.872
MJPEG	0.947	0.891	0.957	0.886
Wavelet	0.873	0.943	0.900	0.921
White noise	0.940	0.943	0.965	0.963
HEVC	0.817	0.943	0.822	0.961
All	0.561	0.717	0.569	0.721

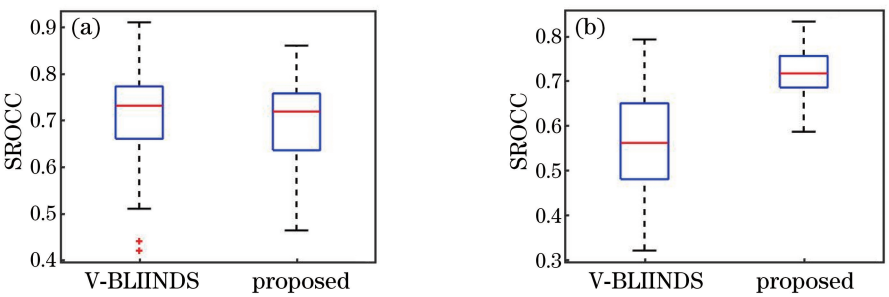

图 3 (a) LIVE 库和(b) CSIQ 库上的 SROCC 分布盒状图

Fig. 3 Box plots of SROCC distributions on (a) LIVE database and (b) CSIQ database

测试视频库后依然能保持较高的评价准确度。如图 3 所示，根据这两种算法在 LIVE 和 CSIQ 视频库上历次测试的 SROCC 结果绘制的盒状图也再次证明了这一点。

4.3 计算复杂度验证

比较本文算法和 V-BLIINDS 算法在质量评价阶段的计算复杂度。由于两者都使用了线性 SVR 根据视频特征预测质量分数，因此实验只记录了两种算法在 LIVE 库和 CSIQ 库上提取失真视频特征所耗费的时间。表 4 列出了这两种算法对视频每一帧进行特征提取平均所耗费的时间。这些实验在同一台计算机上完成，计算机配置为 3.50 GHz CPU 和 12 GB RAM。

通过表 4 可以看出本文算法在质量评价阶段的计算复杂度远低于 V-BLIINDS 算法，这是因为本文算法在评价失真视频时不需要作域变换等复杂度高的运算，因此本文算法能够较为快速地完成对失真视频质量的准确评价，更加符合实际应用的要求。

表 4 每帧特征提取时间(单位：s)

Table 4 Feature extraction time per frame (unit: s)

Database	V-BLIINDS	Proposed
LIVE	23.552	6.572
CSIQ	33.411	8.377

5 结 论

通过在视频质量库上训练三维卷积神经网络模型 3D ConvNets，并在质量评价阶段使用 3D ConvNets 提取失真视频的时空域特征，提出了一种基于三维卷积神经网络的通用型无参考视频质量评价算法。为了评估该算法的性能，在两种不同的视频库上测试验证了算法的评价准确度以及质量评价阶段的计算复杂度。实验结果表明，该算法能够较为准确地评价多种失真类型，在面对不同视频库时保持稳定的评价准确度，稳定性好，且质量评价的计算复杂度显著低于目前主流的无参考视频质量评价算法，评价速度更快。

参 考 文 献

[1] Hou C P, Ma T T, Yue G H, *et al*. Multiply-distorted image quality assessment based on high-order phase congruency[J]. Laser & Optoelectronics Progress, 2017, 54(7): 071001.
侯春萍, 马彤彤, 岳广辉, 等. 基于高阶相位一致性的混合失真图像质量评价[J]. 激光与光电子学进展, 2017, 54(7): 071001.

[2] Zhang Y, Jin W Q. Assessment method of fusion image quality in wavelet domain structural similarity [J]. Chinese Journal of Lasers, 2012, 39(s1):

s109007.
张勇, 金伟其. 小波域结构相似度融合图像质量评价方法[J]. 中国激光, 2012, 39(s1): s109007.

[3] Saad M A, Bovik A C, Charrier C. Blind prediction of natural video quality[J]. IEEE Transactions on Image Processing, 2014, 23(3): 1352-1365.

[4] Li X, Guo Q, Lu X. Spatiotemporal statistics for video quality assessment[J]. IEEE Transactions on Image Processing, 2016, 25(7): 3329-3342.

[5] Negi P S, Labate D. 3-D discrete shearlet transform and video processing[J]. IEEE Transactions on Image Processing, 2012, 21(6): 2944-2954.

[6] Li Y, Po L M, Cheung C H, *et al*. No-reference video quality assessment with 3D shearlet transform and convolutional neural networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(6): 1044-1057.

[7] Wang C F, Su L, Zhang W G, *et al*. No reference video quality assessment based on 3D convolutional neural network[J]. Journal of Software, 2016, 27(2): 103-112.
王春峰, 苏荔, 张维刚, 等. 基于3D卷积神经网络的无参考视频质量评价[J]. 软件学报, 2017, 27(2): 103-112.

[8] Chen J, Jiang H, Liu T D, *et al*. Optimization for Raman fiber amplifiers based on least squares support vector regression model[J]. Acta Optica Sinica, 2015, 35(11): 1123004.
陈静, 江灏, 刘暾东, 等. 基于最小二乘支持向量回归模型的拉曼光纤放大器优化设计[J]. 光学学报, 2015, 35(11): 1123004.

[9] Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision, 2015: 4489-4497.

[10] Seshadrinathan K, Soundararajan R, Bovik A C, *et al*. Study of subjective and objective quality assessment of video[J]. IEEE Transactions on Image Processing, 2010, 19(6): 1427-1441.

[11] Vu P V, Chandler D M. ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices[J]. Journal of Electronic Imaging, 2014, 23(1): 013016.

[12] Wang Z, Bovik A C, Sheikh H R, *et al*. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[13] Seshadrinathan K, Bovik A C. Motion tuned spatio-temporal quality assessment of natural videos[J]. IEEE Transactions on Image Processing, 2010, 19(2): 335-350.

[14] Pinson M H, Wolf S. A new standardized method for objectively measuring video quality[J]. IEEE Transactions on Broadcasting, 2004, 50(3): 312-322.