文章编号:2095-6134(2018)04-0544-06

基于卷积神经网络的时空融合的无参考视频质量评价方法*

王春峰[1],苏 荔[1,2†],黄庆明[1,2]

(1 中国科学院大学大数据挖掘与知识管理重点实验室,北京 100049;
2 中国科学院计算技术研究所智能信息处理重点实验室,北京 100190)

(2017年3月31日收稿;2017年4月25日收修改稿)

Wang C F, Su L, Huang Q M. Spatio-temporal-fused no-reference video quality assessment based on convolutional neural network[J]. Journal of University of Chinese Academy of Sciences, 2018,35(4):544-549.

摘 要 无参考视频质量评价是指在不借助原始无损参考视频信息的条件下,对于给定的任意一段视频,直接评测出其质量程度。传统的无参考视频质量评价方法大都基于统计分析,绝大多数都针对特定的视频失真类型,对视频的时域信息考虑较少,导致现有的基于统计分析的方法应用范围局限,实时性较差。提出一种融合视频时空信息的基于卷积神经网络的无参考视频质量评价方法。该方法不针对特定失真类型。将方法分为空域和时域两部分进行处理,空域上提出一种基于卷积神经网络的方法学习空域失真特征,时域上设计一组基于邻帧块结构相似度的特征用以表征视频的时域失真信息。最后将视频的时空特征进行融合,送至线性回归模型进行视频质量的预测。实验表明,所提方法的多项指标均达到主流视频质量评价方法的性能,且方法运行速度大大提高,显示出较好的实时应用前景。

关键词 视频质量评价;卷积神经网络;无参考;时空信息

中图分类号:T9391.41 **文献标志码**:A **doi**:10.7523/j.issn.2095-6134.2018.04.018

Spatio-temporal-fused no-reference video quality assessment based on convolutional neural network

WANG Chunfeng[1], SU Li[1,2], HUANG Qingming[1,2]

(1 *Key Laboratory of Big Data Mining and Knowledge Management of CAS, University of Chinese Academy of Sciences, Beijing* 100049, *China*; 2 *Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract No-reference video quality assessment (NR-VQA) measures distorted videos quantitatively without the reference of original distorted-less videos. Most conventional NR-VQA methods are based on statistical analysis, and the majority of them are generally designed for specific types of distortions or consider less about the temporal information, which limits their application scenarios as well as their speeds. In this paper, we propose a spatio-temporal no-reference video quality assessment method based on convolutional neural network, which is not designed for specific

* 国家自然科学基金(61650202,61472389,61332016,U1636214)资助

† 通信作者,E-mail:suli@ucas.ac.cn

types of distortions. We divide the method into spatial and temporal processes. We redesign a convolutional neural network in spatiality to learn the distortion features in frames. A group of SSIM-like features are exploited in temporality. Finally, we train a linear regression model using the spatio-temporal features to predict the video quality. Experiments demonstrate that the proposed method is similar to other state-of-the-art no-reference VQA methods in performance. Fourthermore, the proposed method runs much faster than other VQA methods, which makes the proposed method have better application prospects.

Keywords video quality assessment; convolutional neural network; no-reference; spatio-temporal information

随着科技与互联网自媒体的迅速发展,各种多媒体资源出现井喷式增长,视频资源逐渐成为人们工作和生活中极为重要的一种多媒体资源。随着各种终端设备的不断涌现,除传统的个人电脑以及台式机外,智能手机与平板电脑也成为更为便捷的终端接收设备。然而,视频从最初的采集完成到最终呈现至用户的终端设备的过程中,不可避免地会引入一些失真。例如视频压缩过程导致的压缩失真,视频流在信道中传输时产生的丢包失真等。这些都会对用户的最终观看体验造成严重影响。如何有效地衡量视频的失真程度,以求进一步降低失真体验,就显得格外重要。视频质量评价在此背景下应运而生。

视频质量评价指的是人眼对于视频质量变化(通常是下降)的感受程度。视频质量评价在很多视频处理领域都起着重要作用,例如视频的压缩、视频水印、视频重建与增强等。视频质量评价根据评价主体可以划分为主观质量评价和客观质量评价。主观评价的评价主体为人,其评价结果与人的视觉感知最为一致,但是主观评价需要耗费大量的人力和时间成本,且其稳定性不好,因此在实际应用中不是很常用,往往用来辅助设计客观评价方法。客观质量评价方法是由计算机进行打分,评价快速稳定,但是设计与人感知一致的客观算法是非常具有挑战性的。客观方法依据使用原始参考视频信息的多少可以划分为全参考、半参考和无参考 3 种评价方法。代表性的全参考评价方法有 STMAD[1],ViS3[2] 和 MOVIE[3] 等。半参考评价方法有 STRRED[4] 等。全参考与半参考评价方法需要借助原始无损视频的全部或部分信息评价失真视频,这就限制了它们的应用范围。因为在很多场景中,无法获得原始视频,比如在评价数码相机录制的视频时或者传输信道带宽十分有限不足以传输原始视频时。在这些场景下,全参考和半参考的方法都变得不适用,这时只能采用无参考评价方法。无参考评价方法就是不借助任何原始视频信息,直接对任一段视频进行质量评价的方法,其应用范围最广,但是由于参考信息的缺失,它也是 3 种方法中最具挑战性的一种。

现有的大多数无参考评价方法是针对特定视频失真类型设计的,文献[5]提出一种基于拉普拉斯金字塔的无参考视频质量评价方法。该方法分为两步,首先提取视频失真特征,将视频的每一帧分解为多个子带的拉普拉斯金字塔,统计各个子带内部与子带间的信息,将这些统计量作为视频帧的失真特征,在序列级别采取闵可夫斯基池化融合特征。最后将得到的视频特征输入神经网络,预测视频的质量得分。该方法主要适用于视频的压缩失真。文献[6]提出一种基于人眼感兴趣区域的无参考视频质量评价方法。该方法首先通过编码信息获得视频帧中的用户感兴趣区域,然后针对感兴趣区域分析提取模糊失真特征和块效应失真特征,最终将模糊和块效应两种失真与用户感兴趣区域信息进行综合考虑,从而估测出视频的质量。这些方法要评价视频的质量,首先要知道待测视频的失真类型,一旦与目标失真类型不一致,则评价方法将不能使用。因此,为解决这一问题,又有研究者提出不针对特定失真类型的无参考评价方法。文献[7]通过对视频相邻帧的残差图在 DCT 域的统计分析进行建模,包括视频中的运动特征、运动一致性度量和视频抖动特征。并且该方法还将无参考图像评价方法 NIQE[8] 中的特征进行提取,以更好地表示视频的空域特征。最后将这些特征输入至线性回归模型进行视频质量的估计。文献[9]发现视频中相邻帧的残差图的带通滤波系数可以捕捉到视频中的时域失真。它直接从空域对残差帧进行局部统计分析。作者沿水平、垂直、主副对角 4 个方向使用

非对称高斯分布来拟合参数。最后从粗粒度和细粒度 2 种不同尺度综合评价视频质量。

然而,这些方法需要预先进行复杂的统计分析,然后结合人的经验设计出一系列复杂的人工特征。这就限制了它们的泛化能力以及预测准确度,实时性也较差。近两年来,一部分研究者开始着眼于设计基于学习的无参考视频质量评价方法。文献[10]中的方法利用空域字典学习表征帧级特征,然后通过回归模型预测帧级得分,进而采取时域融合策略得到视频得分。文献[11]利用一维卷积神经网络对视频的剪切波特征进行提纯,利用提纯后的特征回归预测视频的质量。

本文在上述方法的基础上,提出一种基于二维卷积神经网络的结合视频时空特征的无参考视频质量评价方法。该方法不针对特定失真类型。将视频分为空域和时域两部分进行处理,最后将视频的时空特征进行融合,送至线性回归模型进行视频质量的预测。

本文的主要贡献如下:

1)提出一种不针对特定失真类型的基于学习的无参考视频质量评价方法,该方法有别于其他基于学习的方法,采用二维卷积神经网络,直接从像素级别学习视频的帧级特征。

2)时域上设计一组新的基于视频邻帧对应块结构相似度的特征。可以有效表示视频时域特征。

3)实验表明,本文所提方法在预测性能上达到主流无参考评价方法的水平,但是其实时速度大大加快,具有很高的实际应用意义。

1 时空融合的无参考视频质量评价

视频与图像有很多相似特性,主要区别在于视频中多了时域特性。因此,本文在设计无参考质量评价算法的过程中,将视频分为空域部分与时域部分分别进行处理。方法的框架流程图如图 1 所示。下面将从 2 个角度出发具体介绍本文的算法细节。

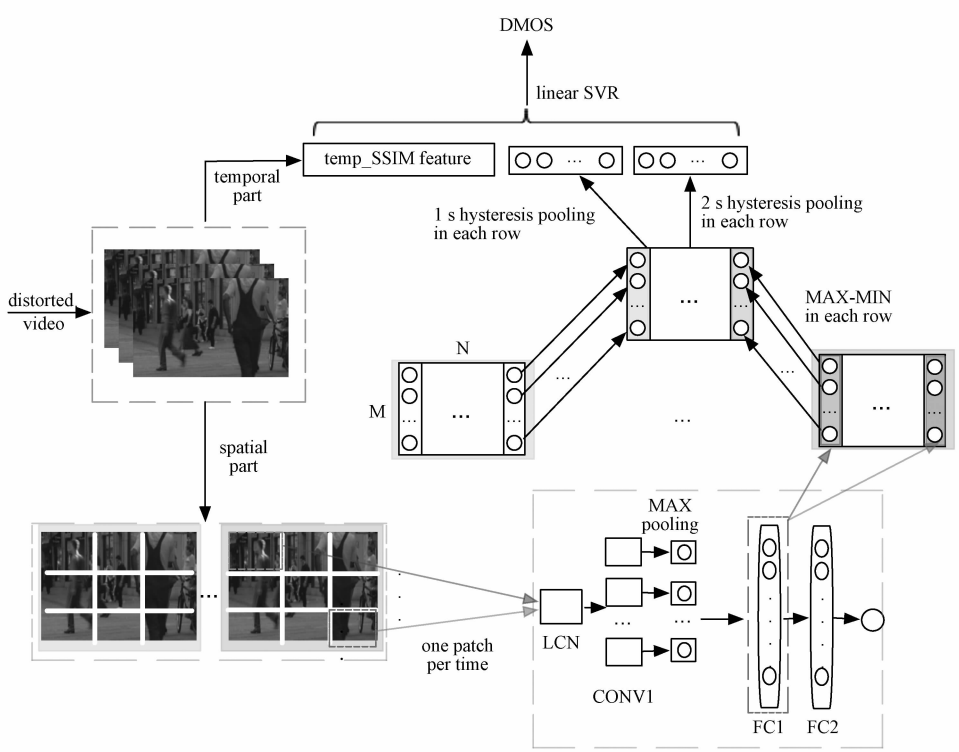

图 1 本文所提方法框架图

Fig. 1 Block diagram of the proposed method

1.1 空域部分

卷积神经网络是基于人类的生物原理设计的,使用卷积神经网络建模失真符合人眼的视觉特性。参照文献[12]中的网络结构,该网络在无参考图像质量评价上取得了不错的效果。该卷积神经网络拥有 1 个卷积层和 2 个全连接层,其中卷积层包括 50 个卷积核且每个卷积核的尺寸为 7×7。2 层全连接层均有 1 024 个节点。最后采

用 L_1 损失函数来训练网络,

$$L = \frac{1}{N}\sum_{n=1}^{N} \| f(x_n, w) - y_n \|_{L_1}, \qquad (1)$$

式中:N 代表样本数量,w 代表网络中的参数,y 表示图像的标注分数,$f(x_n, w)$ 表示网络预测的得分。该网络只设置一层卷积层,主要有 2 个原因:第一是因为标注样本太少,不足以训练更深层的网络;第二是因为考虑到方法的实时性要求,浅层网络可以加快运算速度。具体地,网络的输入为 32×32 的图像小块,每一个小块都预先经过局部对比度归一化的处理,目的是为了减轻饱和度的问题,同时使网络更为鲁棒。整个空域卷积神经网络是在 LIVE 图像评价数据集上进行训练的。

网络训练完成后,需要把小块级别的特征组合为视频级别的特征。本文中选择第一层全连接层的特征作为小块特征。文献[10]发现通过将图像所有小块的每一维特征做最大最小值差值处理可以捕捉质量变化。因此,借鉴此方式将小块级的特征组织为帧级特征。文献[13]提出一种基于磁滞效应的方法可以表征视频的时域质量变化。我们采用此方法将帧级特征组织为视频级特征。在实践中,发现采用 1 s 和 2 s 滞后融合的方式可以进一步提高性能。为验证网络所学特征的有效性,绘制针对不同失真类型和不同失真程度的视频特征统计直方图,如图 2 所示。可以清晰地看到网络所学的特征可以有效区分不同失真等级的视频序列。

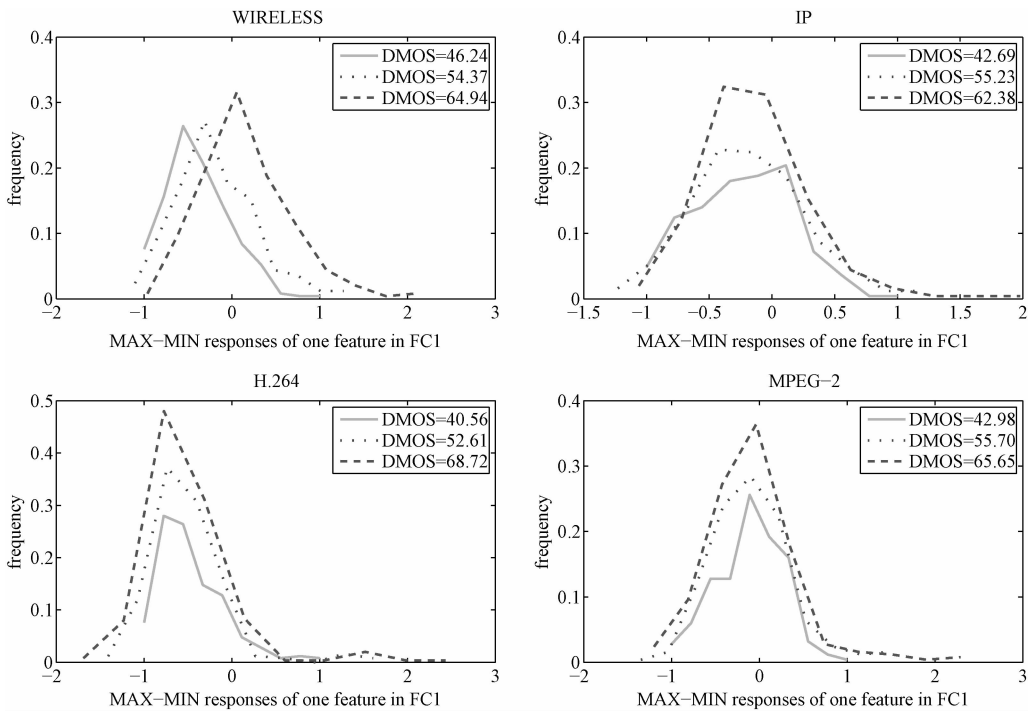

图 2　不同失真类型上不同失真等级的视频特征分布直方图

Fig. 2　Feature distribution histogram for different types and levels of distortions

1.2　时域部分

很多视频失真中都存在局部抖动的现象,视频失真越严重,局部抖动往往越剧烈。可以通过统计分析视频中相邻帧间的变化对这一问题进行建模。将视频帧分为 16×16 的小块,对于每一个小块,采用 3 步搜索算法估计运动向量。我们认为相邻帧间对应小块的结构相似度(SSIM)[14]可以反映视频的时域质量。基于此假设,设计一组时域特征。定义 Variance 表示向量或矩阵的方差,Mean 表示向量或矩阵的均值。假设视频中共有 T 帧,则 SSIM_MAP 矩阵中的每一个元素即为邻帧间对应小块的结构相似度值,然后时域特征表示如下:

$$\mathrm{Mean}_{t=1}^{T-1}(\mathrm{Mean}(\mathrm{SSIM\_MAP}(t))), \qquad (2)$$

$$\mathrm{Variance}_{t=1}^{T-1}(\mathrm{Mean}(\mathrm{SSIM\_MAP}(t))), \qquad (3)$$

$$\mathrm{Mean}_{t=1}^{T-1}(\mathrm{Variance}(\mathrm{SSIM\_MAP}(t))), \qquad (4)$$

$$\mathrm{Variance}_{t=1}^{T-1}(\mathrm{Mean}(\mathrm{SSIM\_MAP}(t))). \qquad (5)$$

其中 t 表示视频中第 t 组邻帧对。

至此,分别得到视频的空域与时域特征。之后将视频的时空特征组合在一起作为视频的最终

特征表示,然后使用该特征训练一个线性支持向量回归模型用来预测视频的质量得分。

2 实验

和多数主流无参考评价方法一样,我们也在 LIVE[15-16] 数据集上对所提方法进行了实验验证。该数据集共有 160 段视频,其中包括 10 段无损视频,每段无损视频又对应 15 段失真类型与失真程度均不同的视频。一共包括 4 种失真类型,分别是 MPEG-2 压缩失真、H.264 压缩失真、IP 网络失真以及无线网络失真。每段视频对应一个主观评分,即不同平均意见得分(different mean opinion score, DMOS),该得分是由数十位受测者进行人为标注的,分值范围为 0 ~ 100 分,分数越高,代表视频的质量越差。

和主流质量评价方法一样,使用 LCC 和 SROCC 两个指标评价算法的性能,其中 LCC 用来度量 2 个变量间线性相关程度,SROCC 主要用来评价算法的预测单调性,两种指标都是值越大表示算法性能越好。训练过程中,将数据集按场景内容划分训练测试集,选取其中 80% 的视频序列作为训练集,剩下 20% 的视频序列作为测试集。考虑所有可能的组合方式,交叉验证 45 次,最后分别选取两种评测指标的中位数作为最终的结果。

首先分别在每种子失真类别上进行实验,之后又在整体数据集上进行实验,并与主流视频质量评价方法进行对比,共对比 4 种全参考评价方法,1 种半参考评价方法和 2 种无参考评价方法。其中 PSNR 和 SSIM[14] 是全参考图像质量评价的方法,本文中先将每一帧计算出得分,最后以所有帧的平均得分作为视频得分。STMAD[1] 和 MOVIE[3] 是目前性能最好的全参考视频质量评价方法。STRRED[4] 是主流的半参考视频质量评价方法,V-BLIINDS[7] 和 VIIDEO[8] 是具有代表性的两种无参考视频质量评价方法。具体的实验对比结果见表 1。

表 1 主流视频质量评价方法在 LIVE 数据集上 SROCC 和 LCC 指标对比

Table 1 SROCC and LCC performance of VQA methods on LIVE

SROCC	方法名称	Wireless	IP	H.264	MPEG-2	ALL
全参考	PSNR	0.691	0.600	0.714	0.643	0.677
	SSIM[14]	0.691	0.543	0.881	0.786	0.650
	MOVIE[3]	0.786	**0.771**	0.881	0.905	0.807
	STMAD[1]	**0.810**	**0.771**	**0.952**	**0.929**	**0.834**
半参考	STRRED[4]	0.762	0.771	0.905	0.905	0.826
无参考	V-BLIINDS[7]	**0.691**	**0.600**	0.643	0.667	**0.735**
	VIIDEO[8]	0.548	**0.600**	**0.762**	0.571	0.651
	本文方法	0.690	**0.600**	0.738	**0.738**	0.671
LCC	方法名称	Wireless	IP	H.264	MPEG-2	ALL
全参考	PSNR	0.798	0.733	0.698	0.696	0.722
	SSIM[14]	0.634	0.726	0.851	0.805	0.625
	MOVIE[3]	**0.920**	0.895	0.919	**0.955**	0.852
	STMAD[1]	0.904	**0.901**	**0.947**	0.942	**0.861**
半参考	STRRED[4]	0.806	0.816	0.892	0.904	0.725
无参考	V-BLIINDS[7]	**0.844**	0.852	**0.956**	**0.949**	**0.790**
	VIIDEO[8]	0.740	0.848	0.886	0.872	0.701
	本文方法	0.808	**0.914**	0.892	0.871	0.713

从表 1 的对比结果可以看出本文所提方法无论在失真子集上还是整个数据集上都达到与主流无参考方法相当的性能,性能甚至要优于经典的全参考评价方法 PSNR 和 SSIM。与 MOVIE、STMAD 和 STRRED 性能差距较大在我们预期内,因为这些方法都用到原始无损视频信息作为参考。与 V-BLLINDS 方法性能有所差距是因为本文方法在空域特征学习时,用到的数据为图像小块,并没有充分的标注信息,这里只是简单地将小块得分粗略等同于图像得分,标注信息存在偏差,影响了最终性能。相信随着标注数据的扩充,所提方法性能将会进一步提高。

我们还与其他无参考方法进一步做了运行速度的对比试验,在 LIVE 数据集上挑选 10 段视频运行算法,以其平均运行时间作为最终结果,具体结果见表 2。

表 2　主流无参考视频质量评价方法在 LIVE 数据集上耗时对比

Table 2　Runtime of NR-VQA methods on LIVE

方法名称	运行时间/s
V-BLIINDS[7]	709.14
VIIDEO[8]	160.94
本文方法	175.63

从表 1 和表 2 可以看出,本文所提方法不仅保证了预测性能,而且运行速度要远远好于 V-BLIINDS 方法,基本与 VIIDEO 方法持平,这使得本文方法的实时性大大提高。

3　总结

本文提出一种融合视频时空特性的基于卷积神经网络的无参考视频质量评价方法。空域上采用二维卷积网络学习空域特征,时域上设计一组基于邻帧对应块结构相似度的特征表征时域失真信息。通过在主流视频质量评价数据集 LIVE 上的实验表明,本文方法达到主流无参考视频评价方法的水平,且运行速度较快,具有较强的实时应用前景。未来还将进一步考虑如何构造监督信息以增强网络在标注数据较少情况下的学习能力。

参考文献

[1] Vu P V, Vu C T, Chandler M D. A spatiotemporal most-apparent-distortion model for video quality assessment[C]// 2011 18th IEEE International Conference on Image Processing. Brussels: IEEE, 2011: 2 505-2 508.

[2] 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): 013 016.

[3] 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.

[4] Seshadrinathan R, Bovik A C. Video quality assessment by reduced reference spatio-temporal entropic differencing[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2013, 23(4): 684-694.

[5] Zhu K F, Keisuke H, Asari V, et al. A no-reference video quality assessment based on laplacian pyramids[C]// 2013 20th IEEE International Conference on Image Processing. Melbourne: IEEE, 2013: 49-53.

[6] Lin X Y, Tian X, Chen Y W. No-reference video quality assessment based on region of interest [C] // 2012 2nd International Conference on Consumer Electronics, Communications and Networks. Yichang: IEEE, 2012: 1 924-1 927.

[7] Saad M A, Bovik A C, Charrier C. Blind prediction of natural video quality [J]. IEEE Transactions on Image Processing, 2014, 23(3): 1 352-1 365.

[8] Mittal A, Soundararajan R, Bovik A C. Making a completely blind image quality analyzer[J]. IEEE Signal processing Letters, 2013, 22(3): 209-212.

[9] Mittal A, Saad M, Bovik A C. Assessment of video naturalness using time-frequency statistics[C]//2014 IEEE International Conference on Image Processing. Paris: IEEE, 2014: 571-574.

[10] Xu J T, Ye P, Liu Y, et al. No-reference video quality assessment via feature learning[C]//2014 IEEE International Conference on Image Processing. Paris: IEEE, 2014: 491-495.

[11] Li Y M, 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): 1 044-1 057.

[12] Kang L, Ye P, Li Y, et al. Convolutional neural network for no reference image quality assessment [C] // 2014 IEEE International Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 1 733-1 740.

[13] Seshadrinathan K, Bovik A C. Temporal hysteresis model of time varying subjective video quality [C] // 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague: IEEE, 2011: 1 153-1 156.

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

[15] Seshadrinathan K, Soundararajan R, Bovik A C, et al. A subjective study to evaluate video quality assessment algorithms[C] // IS&T/SPIE Electronic Imaging. San Jose: IEEE, 2010: 75270H.

[16] Sheikh H R, Bovik A C, Veciana G D. An information fidelity criterion for image quality assessment using natural scene statistics[J]. IEEE Transactions on Image Processing, 2005, 14(12): 2 117-2 128.