



(12) 发明专利

(10) 授权公告号 CN 112784698 B

(45) 授权公告日 2024. 07. 02

(21) 申请号 202011637358.7

(22) 申请日 2020.12.31

(65) 同一申请的已公布的文献号
申请公布号 CN 112784698 A

(43) 申请公布日 2021.05.11

(73) 专利权人 杭州电子科技大学
地址 310018 浙江省杭州市下沙高教园区2号大街

(72) 发明人 殷海兵 刘银豪 周晓飞 王鸿奎

(74) 专利代理机构 杭州浙科专利事务所(普通合伙) 33213
专利代理师 杨小凡

(51) Int. Cl.
G06V 20/40 (2022.01)
G06V 10/82 (2022.01)

G06V 10/44 (2022.01)
G06V 10/52 (2022.01)
G06V 10/80 (2022.01)
G06V 10/764 (2022.01)
G06V 10/766 (2022.01)
G06N 3/0464 (2023.01)
G06N 3/0442 (2023.01)
G06N 3/045 (2023.01)
G06N 3/0985 (2023.01)
G06N 3/048 (2023.01)

(56) 对比文件
CN 109446923 A, 2019.03.08
审查员 王丹丹

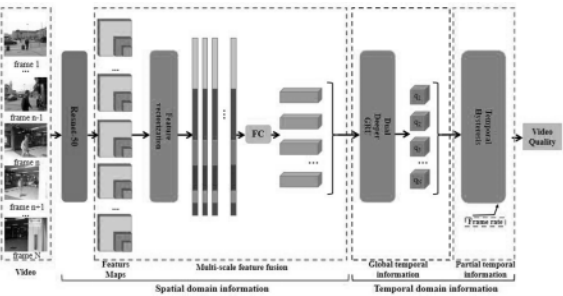
权利要求书2页 说明书9页 附图2页

(54) 发明名称

基于深层次时空信息的无参考视频质量评价方法

(57) 摘要

本发明公开了基于深层次时空信息的无参考视频质量评价方法,包括如下步骤:S1,内容感知特征提取,利用Resnet-50预训练的深度神经网络,提取顶层的语义层特征后进行聚合,对特征图进行均值聚合与标准差聚合;S2,对时间记忆效应进行建模,在特征整合方面,采用GRU网络对长期依赖关系进行建模,在质量聚合方面,提出主观启发的时间池化模型并将其嵌入到网络中;现有的NR-VQA方法无法很好地对VQA任务中的长期依赖关系建模,为了解决这个问题,我们求助于GRU,它是带有门控制的递归神经网络模型,能够集成特征并学习长期依赖关系,本发明中使用GRU集成内容感知功能并预测逐帧质量得分。



1. 基于深层次时空信息的无参考视频质量评价方法,其特征包括如下步骤:

S1,内容感知特征提取,利用预训练的深度神经网络,提取顶层的语义层特征后进行聚合,对特征图进行均值聚合与标准差聚合;

S2,对时间记忆效应进行建模,在特征整合方面,采用GRU网络对长期依赖关系进行建模,在质量聚合方面,提出主观启发的时间池化模型并将其嵌入到网络中;

提取的内容感知特征在输入GRU之前,执行尺寸减少,使用单个全连接层执行尺寸缩减;

$$x_t = W_{fx} f_t + b_{fx}$$

其中 W_{fx} 和 b_{fx} 是该全连接层的参数;降维后,将缩减后的特征 x_t 送入GRU, $t=1, \dots, T$,将GRU的隐藏状态视为集成特征,其初始值为 h_0 ;

根据当前输入 x_t 和先前的隐藏状态 h_{t-1} 计算出的当前隐藏状态 h_t ,即:

$$y_T^f, h_T^f = GRU(x_T^f, h_0^f)$$

$$y_T^b, h_T^b = GRU(y_T^f, h_0^f)$$

$$y_T^b, h_T^b = GRU(x_T^b, h_0^b)$$

$$y_T^f, h_T^b = GRU(y_T^f, h_0^f)$$

最后采用Dual Deeper GRU结构;

$$y_T^b, h_T^b = GRU(x_T^b, h_0^b)$$

$$y_T^f, h_T^f = GRU(y_T^f, h_0^f)$$

$$y_T^f, h_T^f = GRU(x_T^f, h_0^f)$$

$$y_T^f, h_T^b = GRU(y_T^f, h_0^b)$$

$$outputs = y_T^f \oplus y_T^b \oplus y_T^{f'} \oplus y_T^{b'}$$

outputs,即通过Dual Deeper GRU模块之后,得到包含更深层次时域长时间相关特征的帧得分,最后通过全连接层进行回归,得到每帧得分;

利用集成功能outputs,可以通过添加单个全连接层来预测帧质量得分 q_t :

$$q_t = W_{hq}(outputs) + b_{hq}$$

其中, W_{hq} 和 b_{hq} 是权重和偏置参数;

主观启发的时间池化模型,将存储器质量元素定义为先前帧中质量得分的最小值;将当前质量元素定义为下一帧中质量得分的基于排序的加权平均值;近似分数被计算为存储器和当前元素的加权平均值;视频质量计算为近似分数的时间平均池。

2. 如权利要求1所述的于深层次时空信息的无参考视频质量评价方法,其特征包括如下步骤

S1的具体内容,包括如下步骤:

S11,设视频具有T帧,将视频帧 I_t 输入到预训练的CNN模型中, $t=1, 2, \dots, T$,并从其顶部卷积层输出深度语义特征图 M_t :

$$M_t = CNN(I_t)$$

M_t 总共包含C个特征图,然后,对 M_t 的每个特征图应用空间全局聚合,将空间全局平均池

化操作 GP_{mean} 应用于 M_t ,合并空间全局标准偏差 GP_{std} ; GP_{mean} , GP_{std} 操作后输出特征向量分别为 f_t^{mean} , f_t^{std} ;内容感知特征提取中,在使用res5c层特征的基础上,同时提取res4f、res3d层特征,对这些层所提出的特征图采用与res5c特征图相同操作:

$$f_{t-4f}^{mean} = GP_{mean}(M_{t-4f})$$

$$f_{t-4f}^{std} = GP_{std}(M_{t-4f})$$

$$f_{t-3d}^{mean} = GP_{mean}(M_{t-3d})$$

$$f_{t-3d}^{std} = GP_{std}(M_{t-3d})$$

$$f_{t-5c}^{mean} = GP_{mean}(M_{t-5c})$$

$$f_{t-5c}^{std} = GP_{std}(M_{t-5c})$$

S12,将 f_t^{mean} 和 f_t^{std} 串联起来,作为内容感知知觉特征 f_t :

$$f_t = f_{t-5c}^{mean} \oplus f_{t-5c}^{std} \oplus f_{t-4f}^{mean} \oplus f_{t-4f}^{std} \oplus f_{t-3d}^{mean} \oplus f_{t-3d}^{std}$$

其中, $\oplus$ 表示拼接操作。

3.如权利要求1所述的于深层次时空信息的无参考视频质量评价方法,其特征在于将基于排序顺序的权函数替换为可区分权重函数,并将其嵌入网络中,具体地:

将第 t 帧的存储质量元素定义为前几个帧中的最低质量得分:

$$l_t = q_t, t=1$$

$$l_t = \min_{k \in V_{prev}} q_k, t > 1$$

其中 $V_{prev} = \{\max(1, t-\tau), \dots, t-2, t-1\}$ 是所考虑帧的索引集,而 τ 是与时间持续时间有关的超参数;

在第 t 帧构造了当前质量元素 m_t ,使用接下来几帧的加权质量得分,其中分配了更大的权重对于较差质量的帧,具体地,通过可分解的softmin函数,即负线性函数和softmax函数的组合,来确定权重 w_t^k :

$$m_t = \sum_{k \in V_{next}} q_k w_t^k,$$

$$w_t^k = \frac{e^{-q_k}}{\sum_{j \in V_{next}} e^{-q_j}}$$

其中 $V_{next} = \{t, t+1, \dots, \min(t+\tau, T)\}$ 是相关帧的索引集;

τ 与帧率相关,即采用的窗大小与视频帧率相关,使 $\tau = \text{int}(\text{round}(\text{Frame.rate})/2)$;

通过线性组合内存质量和当前质量元素来近似主观帧质量得分,然后,通过近似分数的时间全局平均池GAP计算总体视频质量 Q :

$$q'_t = \gamma l_t + (1-\gamma) m_t$$

$$Q = \frac{1}{T} \sum_{t=1}^T q'_t$$

其中, γ 是一个超参数,用于平衡内存和当前元素对近似分数的贡献。

基于深层次时空信息的无参考视频质量评价方法

技术领域

[0001] 本发明涉及图像、视频处理技术领域,尤其是涉及一种基于深度卷积网络的无参考视频质量评价方法。

背景技术

[0002] 随着具有拍照录像功能的可穿戴设备、智能手机与平板电脑的普及,视频信息的获取和存储变得越来越容易,人们可以随意拍摄不同时长的视频信息,其中不乏大量的失真视频,这极大的影响观众的视觉体验。在整个视频链路中,大部分模块都可以精确度量,如采集、上传、预处理、转码、分发等。然而未知的部分却恰恰是最关键的部分,即用户的视频观看体验到底怎么样。目前行业内的视频质量评估方法分为两大类:客观质量评估与主观质量评估。主观评价方式通过人类视觉系统进行评价,较准确,但不稳定,耗时,且成本较高,难以操作。客观评价方法对人进行建模,通过模型评价图像质量,具有批量处理、结果稳定和可重现的优点。近年来,如何利用计算机技术模拟人类视觉注意力评价机制,正确评价视频质量成为当前研究的热点。在过去的十年中,研究人员提出了众多的评价模型,并应用到很多领域,如图像/视频降噪、图像/视频压缩编码、图像/视频风格迁移等。同时,我们也要注意随着自媒体的传播,拍摄视频存在不确定且混合的失真,而这无疑增加了该领域的研究难度。

[0003] 视频质量评价对象不受限于对象的类别、尺寸大小、失真类型、个数等,这些不确定因素使得它成为目前计算机视觉和人工智能领域中的一个难题。在传统做法中,研究人员根据提取视频中的抽象特征来表示图像。这些特征包括:对比度、亮度、颜色分布、运动光流等,但在复杂的场景中,传统方法往往不够准确。这是因为这些观察往往限于低级别的特征(例如,颜色和对比度等),而不能准确描述人眼对于视频质量的复杂感知。

[0004] 近年来,深度卷积神经网络(Convolutional Neural Network, CNN)广泛运用于计算机视觉的各个领域,很多困难的视觉问题都获得了重大的进展。不同于传统方法,深度卷积神经网络可以从大量的样本中建模并且自动地、端到端地学习到更为本质的特征,从而有效地避免了传统人工建模和设计特征的弊端。在视频质量评价领域,深度卷积神经网络也被广泛地使用,并且大幅度地提高了视频质量评价的准确性和泛化性,但是由于深度网络需要大量运用的池化操作(例如max-pooling和average-pooling)编码上下文信息,这些方法总是不能很好地保存视频中深层次的信息,即得到提取的视频特征仅仅涉及浅层的语义特征与时域相关信息。而事实上,视频中帧内与帧间的深层次相关信息十分重要,我们人眼对视频的感知与评价是一个十分复杂的过程,单凭一个简单的顶层语义与简单的循环卷积神经网络是不够的。同时,忽略这些深层次的时空信息,往往只能得到次优的结果。

[0005] 幸运的是,随着深度学习技术的发展,其已成功应用于许多计算机视觉任务中,例如物体检测、语义分割和视频质量评价领域。最近,研究发现将深度卷积神经网络引入到视频质量评价中,探索时空联合域的深层次时空信息,极大地提升了视频评价的性能。

[0006] 在具体实现中,仍存在如下技术问题:

[0007] 1.内容感知特征提取模块仅仅提取了最顶层的特征。然而在主观实验中,实验对象不仅仅关注最顶层的语义特征,在不同的图像中,所关注的图像特征也不同,图像的特征包括边缘、细节等。这导致评价的时候所得到的信息不够全面。

[0008] 2.对于全局的时序信息,仅仅考虑使用单向的GRU。在主观实验当中,实验对象对于当前帧的质量评价不仅仅受之前帧影响,实际上也受到后续帧的影响。因此,单纯考虑单向的GRU得到的结果是不准确的。

[0009] 3.时间记忆模型中的窗滑动不考虑帧率,只采取固定的窗。时间记忆模型的窗应根据帧率采取不同大小。

发明内容

[0010] 为解决现有技术的不足,实现提升视频评价的性能目的,本发明采用如下的技术方案:

[0011] 基于深层次时空信息的无参考视频质量评价方法,包括如下步骤:

[0012] S1,内容感知特征提取,利用Resnet-50预训练的神经网络,提取顶层的语义层特征后进行聚合,对特征图进行均值聚合与标准差聚合;

[0013] S2,对时间记忆效应进行建模,在特征整合方面,采用GRU网络对长期依赖关系进行建模,在质量聚合方面,提出主观启发的时间池化模型并将其嵌入到网络中。现有的NR-VQA方法无法很好地对VQA任务中的长期依赖关系建模,为了解决这个问题,我们求助于GRU,它是带有门控制的递归神经网络模型,能够集成特征并学习长期依赖关系,本发明中使用GRU集成内容感知功能并预测逐帧质量得分。

[0014] 进一步的,所述步骤S1的具体内容,包括如下步骤:

[0015] S11,设视频具有T帧,将视频帧 I_t ($t=1,2,\dots,T$) 输入到预训练的CNN模型中,并从其顶部卷积层输出深度语义特征图 M_t :

[0016] $M_t = \text{CNN}(I_t)$

[0017] M_t 总共包含C个特征图,然后,对 M_t 的每个特征图应用空间全局聚合,仅仅将空间全局平均池化操作 GP_{mean} 应用于 M_t 会丢弃许多 M_t 信息,因此我们进一步考虑合并空间全局标准偏差 GP_{std} ,以保留 M_t 中的变化信息; $GP_{\text{mean}}, GP_{\text{std}}$ 操作后输出特征向量分别为 $f_t^{\text{mean}}, f_t^{\text{std}}$;内容感知特征提取中,在使用res5c层特征的基础上,同时提取res4f、res3d层特征,对这些层所提出的特征图采用与res5c特征图相同操作:

[0018] $f_{t-4f}^{\text{mean}} = GP_{\text{mean}}(M_{t-4f})$

[0019] $f_{t-4f}^{\text{std}} = GP_{\text{std}}(M_{t-4f})$

[0020] $f_{t-3d}^{\text{mean}} = GP_{\text{mean}}(M_{t-3d})$

[0021] $f_{t-3d}^{\text{std}} = GP_{\text{std}}(M_{t-3d})$

[0022] $f_{t-5c}^{\text{mean}} = GP_{\text{mean}}(M_{t-5c})$

[0023] $f_{t-5c}^{\text{std}} = GP_{\text{std}}(M_{t-5c})$

[0024] S12,将 f_t^{mean} 和 f_t^{std} 串联起来,作为内容感知知觉特征 f_t :

$$[0025] \quad f_t = f_{t-5c}^{\text{mean}} \oplus f_{t-5c}^{\text{std}} \oplus f_{t-4f}^{\text{mean}} \oplus f_{t-4f}^{\text{std}} \oplus f_{t-3d}^{\text{mean}} \oplus f_{t-3d}^{\text{std}}$$

[0026] 其中, $\oplus$ 表示拼接操作。

[0027] 进一步的, 所述步骤S2中, 提取的内容感知特征在输入GRU之前, 执行尺寸减少, 使用单个全连接层执行尺寸缩减, 即4096降维至128维度, 提取的内容感知功能具有较高的维度, 对于训练GRU而言不容易, 因此, 将它们送入GRU之前执行尺寸减小, 通过与优化过程中的其他步骤一起执行降维操作;

$$[0028] \quad x_t = W_{fx} f_t + b_{fx}$$

[0029] 其中 W_{fx} 和 b_{fx} 是该全连接层的参数; 降维后, 将缩减后的特征 x_t ($t=1, \dots, T$) 送入GRU, 将GRU的隐藏状态视为集成特征, 其初始值为 h_0 ;

[0030] 根据当前输入 x_t 和先前的隐藏状态 h_{t-1} 计算出的当前隐藏状态 h_t , 即:

$$[0031] \quad y_T^f, h_T^f = GRU(x_T, h_0^f)$$

$$[0032] \quad y_T^b, h_T^b = GRU(y_T^f, h_0^f)$$

$$[0033] \quad y_T^b, h_T^b = GRU(x_b, h_0^b)$$

$$[0034] \quad y_T^f, h_T^b = GRU(y_T^f, h_0^f)$$

[0035] 最后为充分利用Deeper GRU结构的时间域相关性, 为提高信息交互的数据量尝试采用Dual Deeper GRU结构;

$$[0036] \quad y_T^b, h_T^b = GRU(x_b, h_0)$$

$$[0037] \quad y_T^f, h_T^f = GRU(y_T^f, h_0)$$

$$[0038] \quad y_T^f, h_T^f = GRU(x_b, h_0)$$

$$[0039] \quad y_T^f, h_T^b = GRU(y_T^b, h_0)$$

$$[0040] \quad outputs = y_T^f \oplus y_T^b \oplus y_T^{f'} \oplus y_T^{b'}$$

[0041] outputs, 即通过Dual Deeper GRU模块之后, 得到包含更深层次时域长时间相关特征的帧得分, 最后通过全连接层进行回归, 得到每帧得分;

[0042] 利用集成功能outputs, 可以通过添加单个全连接层来预测帧质量得分 q_t :

$$[0043] \quad q_t = W_{hq}(outputs) + b_{hq}$$

[0044] 其中, W_{hq} 和 b_{hq} 是权重和偏置参数。

[0045] 进一步的, 所述步骤S2中的主观启发的时间池化模型, 将存储器质量元素定义为先前帧中质量得分的最小值; 将当前质量元素定义为下一帧中质量得分的基于排序的加权平均值; 近似分数被计算为存储器和当前元素的加权平均值; 视频质量计算为近似分数的时间平均池。但是, 直接将此模型应用于真实失真视频的质量评估存在一些限制。首先, 该模型需要可靠的帧质量得分作为输入, 这在我们的任务中无法提供。第二, 本申请中的模型是不可区分的, 因为当前质量元素的定义中基于排序顺序的权重, 因此, 它无法嵌入到神经网络中, 由于我们只能访问总体主观视频质量, 因此我们需要学习神经网络而无需帧级监督, 因此, 为了将预测的帧质量得分 q_t 与视频质量 Q 连接起来, 我们提出了一种新的可区分时间池化模型。

[0046] 进一步的,将基于排序顺序的权函数替换为可区分权重函数,并将其嵌入网络中,具体地:

[0047] 为了模仿人类对不良质量事件的容忍度,将第 t 帧的存储质量元素定义为前几个帧中的最低质量得分:

$$[0048] \quad l_t = q_t, t=1$$

$$[0049] \quad l_t = \min_{k \in V_{prev}} q_k, t > 1$$

[0050] 其中 $V_{prev} = \{\max(1, t-\tau), \dots, t-2, t-1\}$ 是所考虑帧的索引集,而 τ 是与时间持续时间有关的超参数;

[0051] 考虑到受试者对质量下降反应敏锐而对质量改善反应迟钝的事实,在第 t 帧构造了当前质量元素 m_t ,使用接下来几帧的加权质量得分,其中分配了更大的权重对于较差质量的帧,具体地,通过可分解的softmin函数,即负线性函数和softmax函数的组合,来确定权重 w_t^k ;

$$[0052] \quad m_t = \sum_{k \in V_{next}} q_k w_t^k,$$

$$[0053] \quad w_t^k = \frac{e^{-q_k}}{\sum_{j \in V_{next}} e^{-q_j}}$$

[0054] 其中 $V_{next} = \{t, t+1, \dots, \min(t+\tau, T)\}$ 是相关帧的索引集;

[0055] 根据实验,尝试不同窗的效果后发现, τ 与帧率相关,即采用的窗大小与视频帧率相关,使 $\tau = \text{int}(\text{round}(\text{Frame.rate})/2)$;

[0056] 通过线性组合内存质量和当前质量元素来近似主观帧质量得分,然后,通过近似分数的时间全局平均池GAP计算总体视频质量 Q :

$$[0057] \quad q'_t = \gamma l_t + (1-\gamma) m_t$$

$$[0058] \quad Q = \frac{1}{T} \sum_{t=1}^T q'_t$$

[0059] 其中, γ 是一个超参数,用于平衡内存和当前元素对近似分数的贡献。

[0060] 本发明的优势和有益效果在于:

[0061] 本发明大幅度地提高了视频质量评价的准确性和泛化性,并且融合了视频中帧内与帧间的深层次相关信息,使得评价时所得到的信息更为全面;对于全局的时序信息,采用双向GRU,使得到的结果更为准确;时间记忆模型中的窗滑动,同时考虑帧率,最终提升了视频评价的性能。

附图说明

[0062] 图1是本发明的网络流程图。

[0063] 图2是本发明中多尺度帧特征融合结构图。

[0064] 图3是本发明中双流更深层次RNN结构图。

[0065] 图4是本发明中基于帧率的自适应平滑局部信息提取模块原理图。

具体实施方式

[0066] 以下结合附图对本发明的具体实施方式进行详细说明。应当理解的是,此处所描述的具体实施方式仅用于说明和解释本发明,并不用于限制本发明。

[0067] 如图1-4所示,基于深层次时空信息的无参考视频质量评价方法的重点,是真实视频的质量评价问题。由于人类是最终用户,因此利用人类视觉系统(HVS)的知识可以帮助建立解决我们问题的客观方法。具体来说,主要是人类对于视频质量的感知分别受到单帧图像内容和短时时间记忆。

[0068] 本发明主要分为如下模块:内容感知特征提取和时间记忆模型。其中内容感知特征提取模块是利用Resnet-50预训练的深度神经网络,提取顶层的语义层特征后进行聚合,对特征图进行均值聚合与标准差聚合。

[0069] 对于真实失真视频,所感知的视频质量在很大程度上取决于视频中视频内容。这可以归因于以下原因:失真的复杂性,人类对于失真的容忍阈值以及人类的偏爱不同的视频内容/场景而异。

[0070] 为了评估真实失真视频的感知质量,上述观察促使我们提取出不仅与失真有关的特征而且对于内容感知相关的特征。我们使用CNN在ImageNet上预训练的图像分类模型,使之具有对于不同内容信息的区分能力。因此,从这些模型中提取的深层功能(例如ResNet)有望实现内容感知。同时,深层特征对失真敏感。因此从预训练的图像分类模型中提取内容感知感知特征是合理的。

[0071] 首先,假设视频具有T帧,我们将视频帧 $I_t (t=1, 2, \dots, T)$ 输入到预训练的CNN模型中,并从其顶部卷积层输出深度语义特征图 M_t :

$$[0072] \quad M_t = \text{CNN}(I_t)$$

[0073] M_t 总共包含C个特征图。然后,我们对 M_t 的每个特征图应用空间全局聚合。仅仅将空间全局平均池化操作(GP_{mean})应用于 M_t 会丢弃许多 M_t 信息,因此我们进一步考虑合并空间全局标准偏差(GP_{std}),以保留 M_t 中的变化信息。 $GP_{\text{mean}}, GP_{\text{std}}$ 操作后输出特征向量分别为 $f_t^{\text{mean}}, f_t^{\text{std}}$ 。

$$[0074] \quad f_{t-4f}^{\text{mean}} = GP_{\text{mean}}(M_{t-4f})$$

$$[0075] \quad f_{t-4f}^{\text{std}} = GP_{\text{std}}(M_{t-4f})$$

$$[0076] \quad f_{t-3d}^{\text{mean}} = GP_{\text{mean}}(M_{t-3d})$$

$$[0077] \quad f_{t-3d}^{\text{std}} = GP_{\text{std}}(M_{t-3d})$$

[0078] 之后,将 f_t^{mean} 和 f_t^{std} 串联起来,作为内容感知知觉特征 f_t :

$$[0079] \quad f_t = f_{t-5c}^{\text{mean}} \oplus f_{t-5c}^{\text{std}} \oplus f_{t-4f}^{\text{mean}} \oplus f_{t-4f}^{\text{std}} \oplus f_{t-3d}^{\text{mean}} \oplus f_{t-3d}^{\text{std}}$$

[0080] 其中, $\oplus$ 表示拼接操作。

[0081] 时间建模是设计目标VQA模型的另一个重要方向。我们从两个方面对时间记忆效应进行建模。在特征整合方面,我们采用GRU网络对我们方法中的长期依赖关系进行建模。在质量聚合方面,本模型提出了一个主观启发的时间池化模型并将其嵌入到网络中。

[0082] 长时间依赖关系建模:现有的NR-VQA方法无法很好地对VQA任务中的长期依赖关系建模。为了解决这个问题,我们求助于GRU。它是带有门控制的递归神经网络模型,能够集

成特征并学习长期依赖关系。具体而言,在本文中,我们考虑使用GRU集成内容感知感知功能并预测逐帧质量得分。

[0083] 提取的内容感知功能具有较高的维度,对于训练GRU而言不容易。因此,最好在将它们送入GRU之前执行尺寸减小。通过与优化过程中的其他步骤一起执行降维操作,可能是有益的。在这方面,我们使用单个全连接层执行尺寸缩减,即4096降维至128维度。

$$[0084] \quad x_t = W_{fx} f_t + b_{fx}$$

[0085] 其中 W_{fx} 和 b_{fx} 是该全连接层的参数。

[0086] 降维后,将缩减后的特征 x_t ($t=1, \dots, T$) 送入GRU模块。我们将GRU的隐藏状态视为集成特征,其初始值为 h_0 。

[0087] 根据当前输入 x_t 和先前的隐藏状态 h_{t-1} 计算出的当前隐藏状态 h_t ,即:

$$[0088] \quad y_T^f, h_T^f = GRU(x_T, h_0^f)$$

$$[0089] \quad y_T^b, h_T^b = GRU(y_T^f, h_0^f)$$

$$[0090] \quad y_T^b, h_T^b = GRU(x_b, h_0^b)$$

$$[0091] \quad y_T^f, h_T^b = GRU(y_T^f, h_0^f)$$

[0092] 最后为充分利用Deeper GRU结构的时间域相关性,为提高信息交互的数据量尝试Dual Deeper GRU结构。

$$[0093] \quad y_T^b, h_T^b = GRU(x_b, h_0)$$

$$[0094] \quad y_T^f, h_T^f = GRU(y_T^f, h_0)$$

$$[0095] \quad y_T^f, h_T^f = GRU(x_b, h_0)$$

$$[0096] \quad y_T^f, h_T^b = GRU(y_T^b, h_0)$$

$$[0097] \quad outputs = y_T^f \oplus y_T^b \oplus y_T^{f'} \oplus y_T^{b'}$$

[0098] 最终的outputs,即通过Dual Deeper GRU模块之后,得到包含更深层次时域长时间相关特征的帧得分。最后通过全连接层进行回归。得到每帧得分。

[0099] 利用集成功能outputs,我们可以通过添加单个全连接层来预测帧质量得分 q_t :

$$[0100] \quad q_t = W_{hq}(outputs) + b_{hq}$$

[0101] 其中, W_{hq} 和 b_{hq} 是权重和偏置参数。

[0102] 双流更深层次RNN,对于全局的时序信息,考虑到后续帧对之前帧的影响。尝试提出双向GRU模块来尝试,正向128维输入得到32维的关联信息,反向输入得到反向的32维关联信息,最后将正反向关联信息concat得到64维向量。

[0103] 同时,为进一步提取全局信息,将前向的GRU进行信息交换,采用串联结构构成一个双向更深层次GRU。如图3所示,正向128维输入得到32维的关联信息,将该32维关联信息反向输入GRU网络得到反向的关联信息。这样构成了正反向关联信息的交互,最后,将正反向关联信息concat得到64维向量。

[0104] 最后,为提高信息交互的数据量,引入后向的GRU,也采用串联结构构成一个Deeper Bidirectional ConvLSTM。首先将反向128维输入得到32维的反向关联信息,再将该32维度反向关联信息再次反向重新输入GRU网络,得到正向的关联信息。这样构成了反正

向关联信息的交互,最后,将反正向关联信息concat得到64维向量。

[0105] 将前后向的conv-GRU得到64维向量进行拼接。即得到全部128维的视频关联信息。

[0106] 主观启发的时间池化模型:得到每帧的得分之后,则需要考虑到时间记忆模型。其具体原理为在主观实验中,实验对象不能容忍质量差的视频事件。在主观实验中发现了时间记忆效应,即受试者对视频质量的下降做出了强烈反应,并在此时间间隔内提供了较差的质量,但对视频质量的改善却做出了钝化的反应。

[0107] 其中,本发明采用了一个时间池模型来说明磁滞效应。具体而言,将存储器质量元素定义为先前帧中质量得分的最小值;将当前质量元素定义为下一帧中质量得分的基于排序的加权平均值;近似分数被计算为存储器和当前元素的加权平均值;视频质量计算为近似分数的时间平均池。但是,直接将此模型应用于真实失真视频的质量评估存在一些限制。首先,该模型需要可靠的帧质量得分作为输入,这在我们的任务中无法提供。第二,论文中的模型是不可区分的,因为当前质量元素的定义中基于排序顺序的权重。因此,它无法嵌入到神经网络中。在我们的问题中,由于我们只能访问总体主观视频质量,因此我们需要学习神经网络而无需帧级监督。因此,为了将预测的帧质量得分 q_t 与视频质量 Q 连接起来,我们提出了一种新的可区分时间池化模型,方法是将中基于排序顺序的权函数替换为可区分权重函数,并将其嵌入网络中。具体如下:

[0108] 为了模仿人类对不良质量事件的容忍度,将第 t 帧的存储质量元素定义为前几个帧中的最低质量得分:

$$[0109] \quad l_t = q_t, t=1$$

$$[0110] \quad l_t = \min_{k \in V_{prev}} q_k, t > 1$$

[0111] 其中 $V_{prev} = \{\max(1, t-\tau), \dots, t-2, t-1\}$ 是所考虑帧的索引集,而 τ 是与时间持续时间有关的超参数。

[0112] 考虑到受试者对质量下降反应敏锐而对质量改善反应迟钝的事实,我们在第 t 帧构造了当前质量元素 m_t ,使用接下来几帧的加权质量得分,其中分配了更大的权重对于较差质量的帧。具体来说,我们通过可分解的softmin函数(负线性函数和softmax函数的组合)来确定权重 w_t^k 。

$$[0113] \quad m_t = \sum_{k \in V_{next}} q_k w_t^k,$$

$$[0114] \quad w_t^k = \frac{e^{-q_k}}{\sum_{j \in V_{next}} e^{-q_j}}$$

[0115] 其中 $V_{next} = \{t, t+1, \dots, \min(t+\tau, T)\}$ 是相关帧的索引集。

[0116] 根据实验发现, τ 与帧率相关,我们尝试 $\tau = \text{int}(\text{round}(\text{Frame.rate})/2)$ 。

[0117] 最后,我们通过线性组合内存质量和当前质量元素来近似主观帧质量得分。然后,通过近似分数的时间全局平均池(GAP)计算总体视频质量 Q :

$$[0118] \quad q'_t = \gamma l_t + (1-\gamma) m_t$$

$$[0119] \quad Q = \frac{1}{T} \sum_{t=1}^T q'_t$$

[0120] 其中, γ 是一个超参数,用于平衡内存和当前元素对近似分数的贡献。

[0121] 利用本发明的方法,在三大公开数据集训练测试,即KoNViD-1k、CVD2014和LIVE-Qualcomm。其中,KoNViD-1k是包含有1200个视频序列的数据集,并提供480参与的人工标注评分结果;CVD2014包含有234个复杂视频序列帧,包含5个不同场景与78种不同拍摄设备;LIVE-Qualcomm包含有208个高质量分辨率的视频帧序列,包含54个场景与8种不同拍摄设备。

[0122] 技术方案的技术效果 (SROCC) 如下表所示:

[0123] 表1 LIVE-Qualcomm

	Method	SROCC↑	KROCC↑	PLCC↑	RMSE↓
[0124]	BRISQUE	0.504(±0.147)	0.365(±0.111)	0.516(±0.127)	10.731(±1.335)
	NIQE	0.463(±0.105)	0.328(±0.088)	0.464(±0.136)	10.858(±1.013)
	CORNLA	0.460(±0.130)	0.324(±0.104)	0.494(±0.133)	10.759(±0.939)
	VIIDEO	0.127(±0.137)	0.082(±0.099)	-0.001(±0.106)	12.308(±0.881)
	VBLINDS	0.566(±0.078)	0.405(±0.074)	0.568(±0.089)	10.760(±1.231)
	VSFA	<u>0.773(±0.061)</u>	<u>0.587(±0.062)</u>	<u>0.795(±0.055)</u>	<u>7.550(±0.700)</u>
	Ours	0.782(±0.039)	0.598(±0.043)	0.816(±0.038)	7.222(±0.470)

[0125] 表2 KoNViD-1k

	Method	SROCC↑	KROCC↑	PLCC↑	RMSE↓
[0126]	BRISQUE	0.654(±0.042)	0.473(±0.034)	0.626(±0.041)	0.507(±0.031)
	NIQE	0.544(±0.040)	0.379(±0.029)	0.546(±0.038)	0.536(±0.010)
	CORNLA	0.610(±0.034)	0.436(±0.029)	0.608(±0.032)	0.509(±0.014)
	VIIDEO	0.298(±0.052)	0.207(±0.035)	0.303(±0.049)	0.610(±0.012)
	VBLINDS	0.695(±0.024)	0.509(±0.020)	0.658(±0.025)	0.483(±0.011)
	VSFA	<u>0.773(±0.019)</u>	<u>0.578(±0.019)</u>	<u>0.775(±0.019)</u>	<u>0.421(±0.021)</u>
	Ours	0.806(±0.015)	0.613(±0.014)	0.815(±0.014)	0.379(±0.011)

[0127] 表3 CVD2014

	Method	SROCC↑	KROCC↑	PLCC↑	RMSE↓
[0128]	BRISQUE	0.709(±0.067)	0.518(±0.060)	0.715(±0.048)	15.197(±1.325)
	NIQE	0.489(±0.091)	0.358(±0.064)	0.593(±0.065)	17.168(±1.318)
	CORNLA	0.614(±0.075)	0.441(±0.058)	0.618(±0.079)	16.871(±1.200)
	VIIDEO	0.023(±0.122)	0.021(±0.081)	-0.025(±0.144)	21.822(±1.152)
	VBLINDS	0.746(±0.056)	0.562(±0.005)	0.753(±0.053)	14.292(±1.413)
	VSFA	<u>0.870(±0.368)</u>	<u>0.695(±0.047)</u>	<u>0.8678(±0.032)</u>	<u>10.8572(±1.352)</u>

[0129]	Ours	0.874(±0.015)	0.702(±0.050)	0.880(±0.032)	10.359(±1.337)
--------	-------------	----------------------	----------------------	----------------------	-----------------------

[0130] 以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述实施例所记载的技术方案进行修改,或者对其中部分或者全部技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明实施例技术方案的范围。

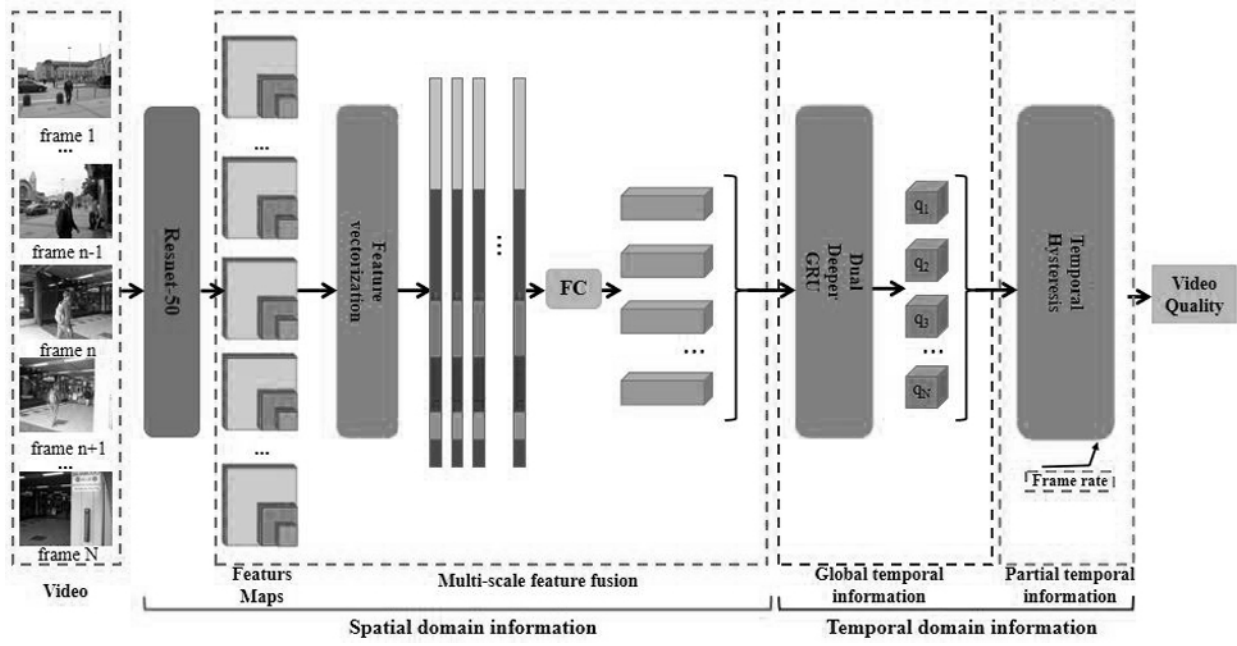


图1

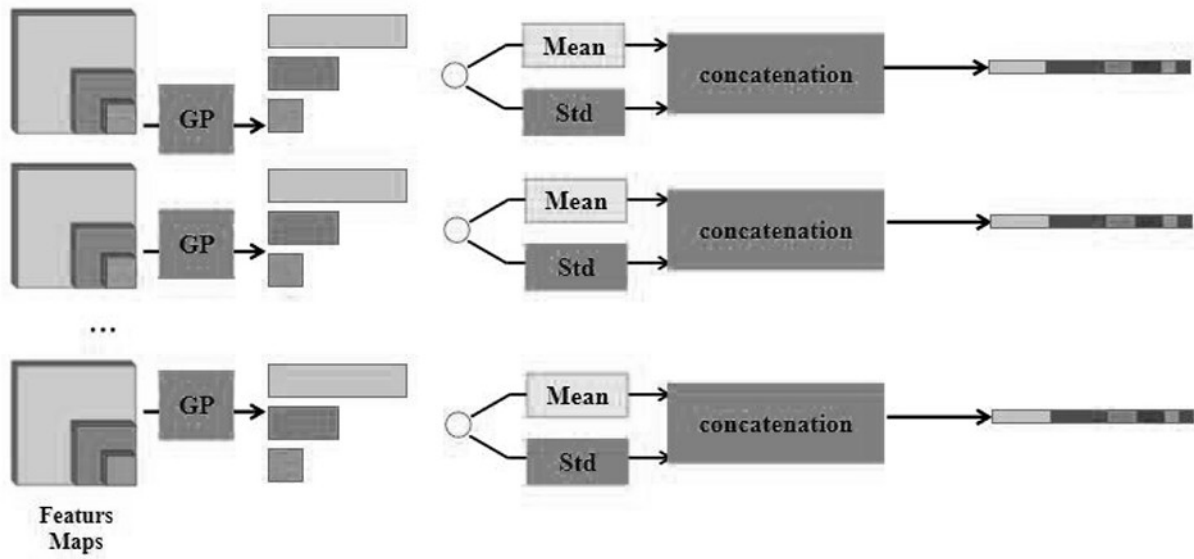


图2

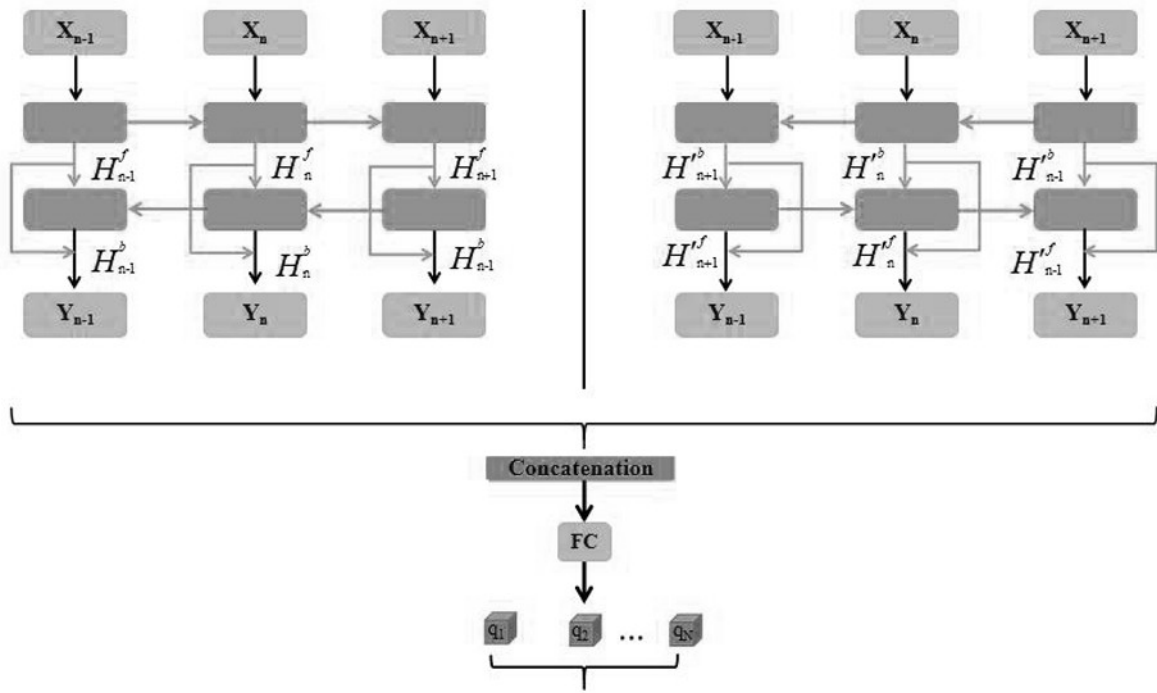


图3

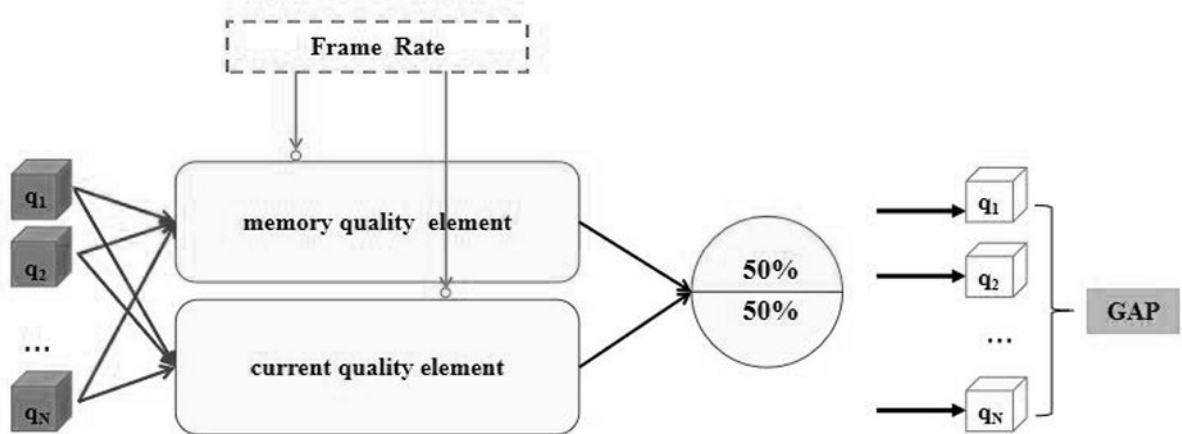


图4