基于视觉注意力变化的网络丢包视频质量评估

冯 欣[1] 杨 丹[2] 张 凌[3]

摘 要 针对网络中受丢包损伤的视频提出了一种基于视觉注意力变化的全参考客观质量评估方法. 该方法基于视觉显著性检测在视频数据上的应用, 考察受网络丢包失真影响的视频数据与标准参考数据在空间和时间上引起的视觉注意力变化, 并根据此变化相应的视觉显著性在空间和时间上的差异, 提出了一组客观质量评估方法. 文中采用 17 个受丢包损伤的视频数据进行测试, 并实施了主观评价实验作为评价标准. 与传统的没有考虑人眼视觉显著特性的质量评估方法, 以及目前主流的基于视觉显著区域/感兴趣区域对失真像素进行加权的方法进行对比, 实验结果表明, 基于视觉注意力变化的方法较后两者与主观质量评估结果有更好的相关性, 能够更有效地评估丢包损伤视频的质量.

关键词 视觉显著性, 注意力变化, 视频质量评估, 网络丢包损伤

DOI 10.3724/SP.J.1004.2011.01322

Saliency Variation Based Quality Assessment for Packet-loss-impaired Videos

FENG Xin[1] YANG Dan[2] ZHANG Ling[3]

Abstract This paper presents a saliency variation based full-reference objective quality assessment metric for packet-loss-impaired videos. The method explores the application of visual saliency information. Motivated by the observation that packet-loss induced errors often change the spatial-temporal visual attention and correspondingly the saliency map, we explore the spatial changes in the saliency values between the original and distorted videos, and the temporal variation of the saliency map of the distorted video. The proposed metric was tested on 17 packet-loss-impaired videos and evaluated by a subjective test. Experiment results showed that the proposed method provides significant improvement in correlation with subjective results over traditional non-saliency quality metrics and the saliency/ROI (region of interest) weighted pixel-error quality measurements. This demonstrates that saliency variation is effective in evaluating the perceptual quality of videos affected by packet loss induces errors.

Key words Saliency, attention change, video quality assessment, packet loss distortion

网络视频随着宽带和无线网络的迅速发展已日渐成为目前网络的主要承载内容. 同时, 不断推陈出新的高品质视频技术使得终端用户对高质量视频的需求越来越大. 然而, 由于存储容量和带宽限制, 视频在传输过程中会遭遇信道传输的拥塞或延迟造成的数据包丢失, 从而影响解码后终端视频的质量. 因此, 网络视频会具有由丢包带来的马赛克现象、局部变形 (图像的某些区域不连续、不清晰)、屏幕局部频繁刷新或闪烁、图像静止等损伤. 而这些失真现象会给终端用户带来明显的视觉影响, 甚至会使网络视频的优势大打折扣. 因此, 对受丢包影响的视频数据进行有效的评估, 特别是建立一种符合人眼视觉感知特性的客观评估方法, 对于网络服务的设计和监控具有重要的意义.

网络视频的编解码技术采用了空间–时间的运动估计机制, 这使得解码后受丢包影响的视频具有独特的感知视觉特征. 例如: 一个丢包带来的错误可能出现在图像的正中间或一个运动的场景中, 而容易引起人们的注意; 也可能出现在背景、角落或者相对静止的区域而被人们忽略. 传统的客观质量评估方法 (如: 峰值信噪比 PSNR、平方差 MSE) 是对所有失真像素进行平均考虑, 却忽略了不同区域的失真会给用户带来不同的视觉感受. 尤其是对于具有明显空间局部性的丢包失真, 这类方法的不足更加明显. 目前大多数融入人类视觉特性的质量评估方法都主要关注于由视频压缩引入的编码失真. 其中, 应用较成功的有 Wang 等提出的基于图像结构相似性度量 (Structural similarity, SSIM)[1] 的评估方法, 该方法认为人眼对场景中的结构信息较敏感, 因此, 通过分别计算图像的亮度、对比度与结构信息之间的相似性, 可以计算出失真图像的质量. 该方法的评估性能较传统的 MSE, PSNR 等

收稿日期 2010-06-12 录用日期 2011-05-21
Manuscript received June 12, 2010; accepted May 21, 2011
国家自然科学基金 (60975015), 中央高校基本科研业务费资助项目 (CDJRC10160010), 教育部博士点基金 (20090191110023), 重庆市自然科学基金 (2010BB2242) 资助
Supported by National Natural Science Foundation of China (60975015), Fundamental Research Funds for the Central University (CDJRC10160010), Doctoral Foundation of Ministry of Education of China (20090191110023), and Natural Science Foundation of Chongqing (2010BB2242)
1. 重庆理工大学计算机科学与工程学院 重庆 400050 2. 重庆大学软件工程学院 重庆 400030 3. 重庆通信学院 重庆 400035
1. College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400050 2. School of Software Engineering, Chongqing University, Chongqing 400030 3. Chongqing Communication Institute, Chongqing 400035

有很大提高. 近几年, 网络视频的迅速发展使得面向网络丢包失真视频的质量评估研究得到了越来越多的关注. 这些方法主要是从网络视频业务的服务质量 (Quality of service, QoS) 和用户体验质量 (Quality of experience, QoE) 两个方面来展开研究. 基于 QoS 的网络视频质量评估方法主要从网络技术和性能的角度分析, 通过探寻影响网络视频质量的网络因素来评估网络视频的损伤情况[2−4]. 基于 QoE 的网络丢包视频质量评估方法则主要从分析解码图像或视频的特征来评估受损图像和视频的质量. 例如, 文献 [5−6] 通过分析受丢包损伤的 MPEG 视频帧的图像特征 (长度, 强度) 提出了一种无参考质量评估方法. Kanumuri 等对视频序列中丢包的可见性建模方面做了大量的工作[7−8]. 文献 [9] 通过对影响丢包视频的几个重要的因素 (错误发生的位置、长度以及受损程度等) 的分析, 提出了一种基于 PSNR 的客观质量评估方法. 专门从事图像/视频质量评估的 LIVE 工作组 (Laboratory for Image and Video Engineering) 最近发布了一个包含丢包损伤的视频数据库[10], 并将目前主流的图像/视频评估方法在该数据集上进行了测试[11]. You 等在文献中将目前的一些客观质量评估方法应用于网络丢包损伤视频, 并对这些方法进行了综合对比评估[12]. 本文也将从用户对网络视频感知质量满意度的角度出发, 分析受网络丢包损伤的解码视频的视觉特征并提出融入视觉显著注意特性的客观质量评估方法.

生理和心理研究表明, 人类总是主动地特别关注于某些特定的、能够产生新异刺激的区域, 这些区域被称为注意焦点 (Focus of attention, FOA) 或显著 (Saliency) 区域. 近年来, 基于视觉注意模型 (Visual attention model) 的图像/视频客观质量评估方法成为了一个新的研究热点[13−15]. 这些方法通过提取图像的感兴趣区域或显著区域来构造 “重要图 (Importance/significant map)” 或者 “显著图 (Saliency map)”, 并以此对检测到的图像/视频失真进行加权. 实验结果表明这些方法都比没有考虑视觉注意因素的方法有不同程度的改进, 但它们主要局限于受编码失真 (如 JPEG, JPEG2000) 的图像或视频. 在失真的视觉效果上, 图像的编码失真表现为整个视野范围内统一分布的量化噪声, 视觉注意区域主要由图像本身的场景内容决定, 而受丢包损伤的视频帧在空间上则表现出局部的不连续、局部显著的异常错误等特征, 这些错误在视觉上更容易引起观看者的注意, 因此, 采用视觉显著性注意机制的质量评估方法更适合于受丢包失真影响的视频.

本文将利用人眼的显著注意特性对受丢包损伤的视频提出一种基于图像失真引起的空间和时间视觉注意力变化的质量评估方法. 视觉显著性检测采用 Itti 提出的自底向上视觉显著性注意检测模型 (Saliency-based bottom-up visual attention model, I-SVAM), 并在该模型上增加运动显著特征检测. 实验数据采用含有单个丢包错误的视频数据, 并按照 ITU-R BT500-11[16] 标准实施了一个主观评估实验作为评价标准.

1 融入显著运动特征的视觉显著区域检测

显著区域检测是通过模拟人眼视觉选择性注意的过程来对视觉注意区域的检测构建客观的可计算模型. 近年来, 对显著区域检测模型的研究已经取得了很多成果, 其中, 最典型的是 Itti 等[17] 提出的基于显著视觉注意力的自底向上显著区域检测模型 (I-SVAM), 该模型已广泛应用于图像/视频压缩、机器人视觉等领域.

1.1 I-SVAM 概述

I-SVAM 检测图像中能够自动地、显著地、无意识地吸引人们视觉注意的区域. 模型首先通过线性滤波将输入图像分解成多个多尺度的低级视觉特征 (灰度、颜色、方向) 通道 (高斯金字塔), 其中, 颜色特征计算红绿 – 蓝黄对抗色, 方向信息通过 Gabor 方向滤波得到 4 个方向特征 ($\theta \in \{0°, 45°, 90°, 135°\}$). 对于每个特征空间通道, 利用中央刺激 – 周围抑制策略 (中心 – 周围尺度差分) 计算视觉感受野, 并应用一种非线性空间竞争机制将不同尺度信息进行合并得到特征显著图 (Conspicuity map). 随后, 通过使用预设的权值对各特征显著图进行线性合并得到一张显著图 (Saliency map). 最后, 神经网络的赢者全胜 (Winner-take-all) 和抑制返回 (Inhibit-of-return) 机制相互作用, 并按照显著级别从高到低产生人们的注意焦点 (FOA). 本文使用 Bernhardt-Walther 实现的 I-SVAM 的 Matlab 工具箱 Saliency Toolbox 1.0[18] 得到视频的注意焦点.

1.2 I-SVAM 中运动显著特征检测

运动特征在视频内容的感知中起着重要的作用, 尤其是在对视频时间域的失真捕捉上. 由于 Saliency Toolbox 1.0 本身没有考虑运动信息, 本文将运动作为另一种视觉显著特征, 在 I-SVAM 中实现一种基于生物相关的多尺度运动感知模型. 遵循 I-SVAM 中显著特征检测的基本思想, 能够引起视觉注意的运动显著特征应是一个区域的运动属性相对于其周围区域运动的显著变化. 因此, 本文利用 I-SVAM 中的多尺度特征描述, 在高斯金字塔的每一层, 应用 Hassenstein-Reichardt (HR) 生物相关运动检测模型[19]. HR 运动检测模型基于对甲虫

视动反应的研究, 通过目标亮度与周围的差异在时间–空间上的相关性来检测运动感知 (见图 1).

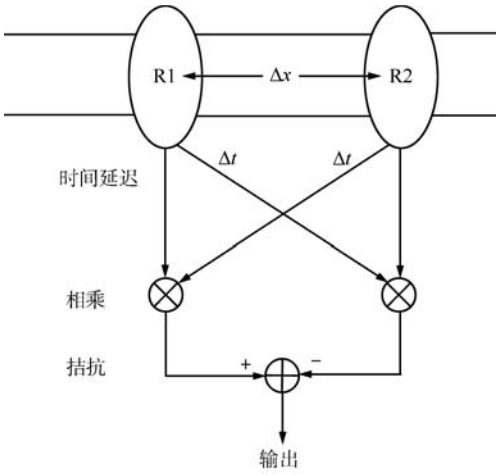


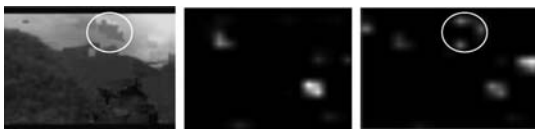
图 1 Hassenstein-Reichardt (HR) 生物相关运动检测模型

Fig. 1 Hassenstein-Reichardt (HR) correlation based motion perception model

在 I-SVAM 中, 对于其构建的灰度金字塔的每一层 σ ($\sigma \in \{0, 1, \cdots, 8\}$), 每相邻两帧应用 HR 运动检测 (对应图 1 中 Δt 为一帧间隔), 并分别在上 (↑)、下 (↓)、左 (←)、右 (→) 4 个方向上检测以 1 像素/帧 ($\Delta x = 1$) 的速度在该方向的运动情况, 从而得到 4 个方向的运动特征图. 在高斯金字塔的二进下采样机制下, 第 σ 层中 1 个像素在某个方向的运动等同于考察原图像 (第 0 层) 以 2^{σ} 的速度在此方向运动的情况, 因此, 所构造的运动金字塔在 4 个方向上检测运动变化幅度范围为 $[0, 2^{2+\sigma}]$. 与计算其他特征的特征显著图 (Conspicuity map) 类似, 利用 I-SVAM 中的 "中央刺激周围抑制" 机制来构造视觉感知的局部运动变化, 并将各尺度各方向的运动差异图合并产生运动特征显著图:

$$\overline{M} = \sum_{d\in\{\leftarrow,\rightarrow,\uparrow,\downarrow\}} \mathcal{N}\left(\bigoplus_{c=2}^{4}\bigoplus_{s=c+3}^{c+4}\mathcal{N}(M(c,s,d))\right) \tag{1}$$

其中, $M(c, s, d)$ 表示在方向 d ($d \in \{\leftarrow, \rightarrow, \uparrow, \downarrow\}$) 上, 中心尺度 c 与周围尺度 s 之间的运动差异, $\mathcal{N}$ 是非线性规则化算子, 在不断迭代中实现局部和周围显著区域的竞争演化, 不同的迭代次数将产生不同大小的显著区域. "$\bigoplus$" 表示跨尺度相加运算. 由于显著运动是视频质量评估的重要因素, 因此, 在对各特征显著图进行加权线性合并得到显著图 (Saliency map) 时, 考虑将运动特征分配较高的权重. 经过对各种权重值的实验, 以及对非线性规则化算子的不同迭代次数分析, 本文发现对于受丢包影响的视频序列, 将灰度、颜色、方向和运动特征的权重值分别设置为: 0.3, 0.3, 0.7, 1.0, 迭代次数为 1, 能够得到与人眼视觉匹配较好的显著图. 图 2 以 "Aircraft" 和 "Leaf" 受丢包损伤的视频序列帧为例, 对在 I-SVAM 中加入显著运动特征与没有考虑运动信息的检测结果进行对比. 如图 2 中椭圆标记处, 在 "Aircraft" 中可以明显看到运动信息检测到了出现在飞机背景中的明显错误, 而在质量较好的 "Leaf" 中也检测到了随微风摆动的树叶. 由此可见, 融入显著运动特征的 I-SVAM 更符合人眼对丢包损伤视频的关注结果.



(a) 示例视频帧 "Aircraft"

(a) Example video frame "Aircraft"

(b) 示例视频帧 "Leaf"

(b) Example video frame "Leaf"

图 2 融入显著运动特征与没有考虑运动信息的 I-SVAM 检测结果对比. (其中, 每一示例视频包括的子图从左到右依次为: 受损视频帧 (其中, "Aircraft", MOS = 2.45 选自序列的第 11 帧; "leaf", MOS = 5, 选自序列的第 12 帧); 原 I-SVAM 检测的显著图; 加入显著运动特征的 I-SVAM 检测得到的显著图.)

Fig. 2 Comparison of I-SVAM with motion saliency feature and without motion (Included in each subimage are: packet-loss-impaired video frame ("Aircraft", MOS = 2.45, the 11th frame is shown; "Leaf", MOS = 5, the 12th frame is shown); saliency map detected by original I-SVAM; saliency map detected by I-SVAM with motion saliency feature.)

2 实验数据及主观质量评估方法

主观评估方法是能够反映人眼观看结果最准确的评价方法, 但其复杂的实施过程、大量人力、物力的花费代价, 使得它不适合于视频的实时质量评估, 而常被作为客观质量评估方法性能评价的标准. 本文采用一组标准视频序列, 通过模拟单个网络丢包事件构造了网络丢包失真视频数据集. 同时, 遵循 ITU-R BT500-11[16] 标准严格实施了一个主观评估实验, 其结果为客观质量评估方法提供评价标准.

2.1 实验数据集

实验采用 17 个分辨率为 320 像素 × 240 像素、帧率为 12 fps 的视频序列作为原始视频序列. 17 个

视频序列均为来自美国某视频研究所的标准数据集, 其中包含了各种室内、室外的自然场景, 不同拍摄状态和运动程度. 由于显著区域检测没有考虑人脸特征, 因此, 本文测试数据的选择侧重于前景图像中没有人脸的视频. 17 个测试数据的场景描述如表 1 (其中训练数据只为使观测者适应实验环境, 其评测结果不作为正式的实验数据).

表 1 测试视频数据集描述 (前 5 个为训练数据集)

Table 1 Descriptions of original video dataset (the first 5 are the training data)

序号	序列名称	序列场景描述
1	F1	一辆赛车在跑道上奔驰
2	Car	一辆轿车飞驰而过
3	Bottles	待处理的瓶子在生产线上缓慢移动
4	Wave	湖面水波荡漾
5	Plane	一架飞机在空中飞过, 离视线较远处
6	Bus	一辆公共汽车穿过闹市区
7	Leaf	树叶在微风中摆动, 摄像机无移动
8	Optis	一些帆船在海面上漂浮
9	Ship	在海面上缓慢前行的游船
10	Stockholm	城市风景鸟瞰
11	Whale	海豚表演, 背景有许多观众
12	Living room	摄像机移动, 餐厅家具及装饰
13	Liberty	自由女神像, 一艘游船缓慢驶过
14	Mobile	玩具火车向左行使和日历上下移动
15	Bedroom	摄像机移动, 起居室床及桌子
16	Boat	一艘小船从海面上驶过, 一个男人站在船上
17	Aircraft	一架直升飞机快速飞来, 离视线越来越近

基于原始视频序列, 本文模拟实现视频序列经过 “编码 − 传输 − 遭受网络丢包 − 解码” 的全过程. 由于长时间丢包失真视频序列的质量需要同时考虑丢包的位置、个数、错误的长度以及宽容效应 (Forgiveness effect) 等因素的影响, 因此, 为了探寻能够准确反映视觉感知质量的显著注意信息, 本文只模拟构建含有单个丢包事件的短视频序列.

从每一个原始视频序列中截取时间为 2 ∼ 4 秒、具有独立完整视频场景的视频段, 并编码成结构为 IPPP $\cdots$ 的一个 GOP (Group of pictures). 实验统计发现视频序列长度为 2 秒就足够观看者对视频给出合理的判断, 并且在这段时间的错误蔓延能够造成合理范围内的质量受损. 对每一个截取的短视频序列, 采用 JM10.0[20] 编码/解码器根据 H.264 基本框架 (Baseline profile) 进行编码/解码, 其参数设置同样为: 目标码率为 128 kbps, 量化系数 $QP = 31$, 平均 PSNR = 34 dB. 视频的每一帧被整个编码成一个 Slice, 并作为一个 RTP (Real-time transport protocol) 包封装传输. 为了排除丢包发生位置的影响, 以及让丢包发生的位置造成正常范围内的失真, 在模拟网络丢包时, 本文选择统一丢弃视频序列的第 3 帧、第 4 帧, 从而模拟传输链路层中两个连续数据包的丢失, 或者由于网络拥塞造成不止一个数据包的丢失. 这个初始丢包错误将在空间和时间上造成连续的错误蔓延, 直到视频的结束. 解码后丢失的视频帧采用帧复制的错误掩盖方式[21], 即被丢弃帧的补偿由上一个正确接收到的视频帧复制得到.

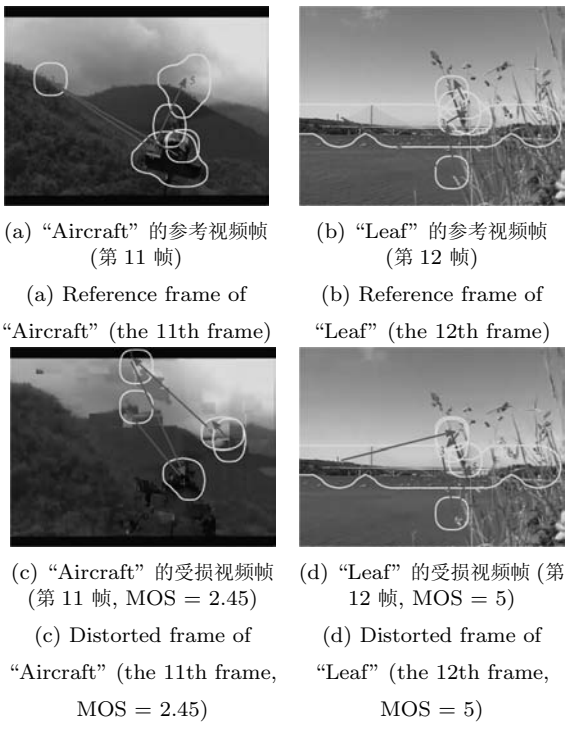
2.2 主观评估实验

主观质量评价仍然采用单激励质量度量 (Single stimulus methods, SSM), 即每一观测者只观看受损视频, 并被告知所观看视频都有不同程度的受损情况. 观测者在观看一遍之后在 1 (最差) ∼ 5 (最好) 的整数范围内给出合理的分数. 为了使观看者适应实验环境, 每一观看者需首先观看 5 个训练视频 (表 1 中前 5 个视频), 这 5 个序列包含了质量受损的全范围, 剩下的 12 个视频随机给出. 32 位非专家观测者对每一个测试视频序列进行评估打分, 无间隔休息地观测所有视频序列. 对一个视频完成观测平均约需要 2 分钟. 观测设备是 17″ LCD 显示器, 观测者可以自由地调节观测距离和角度, 通常观测者会选择 4 ∼ 6 倍图像大小的距离. 实验其他方面的参数遵循 ITU-R BT500-11[16] 推荐设置. 32 位观测者为一个视频序列打分的平均值即为该视频序列的主观平均评定得分 (Mean opinion score, MOS), 这一结果为本文提出的客观质量评估方法提供参考标准.

3 基于视觉注意力变化的视频质量评估方法

人眼的视觉活动是一个极其复杂的过程, 它既包括自底向上数据驱动的初级视觉预注意阶段, 也包括自顶向下由大脑感知控制的高级视觉感知阶段. 视觉的高级注意阶段是在来自大脑皮层的认知和感知信息的指导下, 对视觉注意场景进行有意识的注意阶段, 因此, 根据每个人知识结构和任务驱动的不同, 所产生的视觉注意区域也不尽相同. 然而, 与个体的知识结构和目标任务无关的是: 人们在长期自然场景的生活和观察中积累的对正常状态、质量无损自然场景 (包括自然景色、人工物体和场景) 的视觉感知经验, 例如每个人的经验里都有着对自然物体的形状、大小、颜色、明度和位置关系的度量. 因此, 当这些隐性的先验知识与其他显性的知识系统或搜索任务一起以自顶向下的方式驱动视觉有意识地注意时, 将根据场景的不同出现以下两种情况: 1) 如果场景的画面正常, 那么视觉注意主要以显性的知识为导引, 或是完成一次有目的的搜索任务; 2) 如果画面发生了不正常的失真或变形, 那

么由于这些失真与隐性的先验感知信息产生了差异, 这些隐性的经验在此时会变成显性的驱动, 使视觉注意转向与先验感知产生差异的失真区域. 同时, 两者之间的差异越大, 越能引起视觉有意识的注视. 如图 3 的示例视频帧 “Aircraft” 中, 在画面正常、质量无损的原参考帧中 (图 3 (a)), 飞翔的直升飞机是画面中的显著对象, 这种情况下, 人眼会在自己的兴趣和观看目的的驱动下追踪飞机的移动; 然而, 当图像的背景区域 “山峰”、“天空” 中出现一些异常不连续、不平滑的区域 (受损视频帧图 3 (c) 中由丢包引入的错误匹配区域), 由于与先验感知期望的注意区域产生了差异, 人眼会被这些异常所吸引, 并将注意力从直升飞机转移到这些区域上. 但在 “Leaf” 中, 由于丢包错误并没有造成明显的视觉影响, 其相应的参考帧 (图 3 (b)) 与受损帧 (图 3 (d)) 注意焦点 (图 3 中圆圈区域) 的位置变化也不大. 而对应它们的主观评估结果, “Aircraft” 的分数较差 (MOS = 2.45), “Leaf” 的分数较高 (MOS = 5). 因此, 视觉注意力从参考视频到受损视频空间变化的多少能够反映受丢包影响视频的质量. 通过研究参考视频和受损视频视觉显著区域的空间–时间差异, 本文提出了一组基于视觉显著注意变化的质量评估方法.



(a) “Aircraft” 的参考视频帧 (第 11 帧)

(a) Reference frame of “Aircraft” (the 11th frame)

(b) “Leaf” 的参考视频帧 (第 12 帧)

(b) Reference frame of “Leaf” (the 12th frame)

(c) “Aircraft” 的受损视频帧 (第 11 帧, MOS = 2.45)

(c) Distorted frame of “Aircraft” (the 11th frame, MOS = 2.45)

(d) “Leaf” 的受损视频帧 (第 12 帧, MOS = 5)

(d) Distorted frame of “Leaf” (the 12th frame, MOS = 5)

图 3 参考视频帧与丢包失真视频帧视觉注意区域对比. (边缘曲线代表 I-SVAM 检测出来的前 5 步 FOA)

Fig. 3 Comparison of I-SVAM detected FOA of packet-loss-impaired frames and their corresponding reference frames (The boundary shows the detected first 5 FOAs.)

3.1 视频空间域显著注意变化

由于网络丢包视频在空间上具有局部不连续、局部明显异常的视觉特征, 因此, 从视觉的注意力角度分析, 如果与原参考视觉场景的期望注意区域相比, 受损视觉场景的注意焦点在空间上发生了变化, 那么这些局部显著的丢包错误是引起视觉注意力转移的主要刺激因素, 而视觉注意力在空间上的变化程度也在一定程度上反映了视频的感知失真. 因此, 本文提出通过描述视觉注意力空间域的变化, 即参考视频与受损视频之间的空间显著性差异 (Saliency deviation, SD) 来度量丢包失真视频的感知质量.

令 $\hat{S}(x,y,z)$, $S(x,y,z)$ 分别表示原参考帧与相应受损帧在像素 (x,y), t 时刻的显著值, 其中参考视频与受损视频空间显著性差异的计算分别从均方差、平均绝对误差和结构差异度三个方面考虑. 最后, 根据不同方法检测得到的显著注意变化对整个视频空间的联合失真进行统计, 并得到视频感知质量的客观评价结果. 其计算公式如下:

1) 基于均方差 (Mean square error, MSE):

$$\mathrm{SDMSE} = M_t(E_{(x,y)}(|\hat{S}(x,y,t) - S(x,y,t)|^2)) \tag{2}$$

其中, $E_{(x,y)}$ 是图像域内的平均值算子, M_t 是整个视频在时间域上的平均算子.

2) 基于平均绝对误差 (Mean absolute difference, MAD):

$$\mathrm{SDMAD} = M_t(E_{(x,y)}(|\hat{S}(x,y,t) - S(x,y,t)|)) \tag{3}$$

3) 基于结构差异度 (Difference structural similarity, DSSIM)

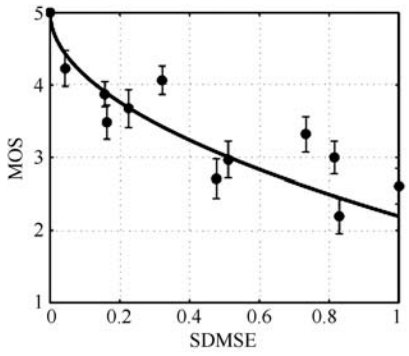
结构差异度计算是基于结构相似度的质量度量方法 (SSIM). 根据文献 [1] 中 SSIM 的计算, 对于参考帧 $\hat{S}$ 以及受损帧 S 中每一像素 (x,y,t) 分别计算其亮度函数、对比度函数和结构函数, 并得到 $\hat{S}$ 和受损帧 S 的结构相似图 $\mathrm{SSIM}_{\mathrm{map}}$. 这里, 由于 SSIM 是计算两幅图像的正相似度, 其取值范围为 $[-1, 1]$, 而其中负值是极少出现的情况, 因此, 为了计算的稳定性以及统一与其他差异评估方法比较, 本文舍弃 SSIM 中的极少数负值且有关 SSIM 计算统一使用结构差异度 (DSSIM) 替代原 SSIM, 由此 SDDSSIM 的计算定义为

$$\mathrm{SDDSSIM} = M_t(\mathrm{DSSIM}(\hat{S}(x,y,t), S(x,y,t))) \tag{4}$$

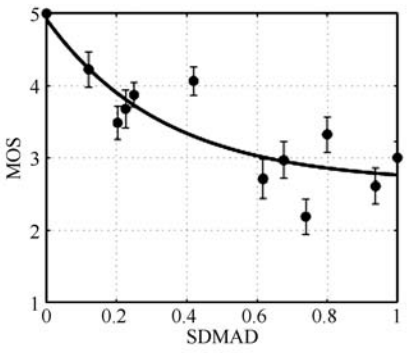
其中, DSSIM 的计算公式为

$$\mathrm{DSSIM}(\hat{S}(x,y,t), S(x,y,t)) = 1 - \max(0, \mathrm{SSIM}_{\mathrm{map}}(\hat{S}(x,y,t), S(x,y,t))) \tag{5}$$

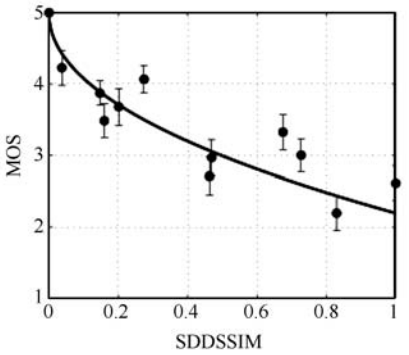
从图 4 的散点图可以看出, 以上三种基于 SD 的评估方法与 MOS 都有较好的负相关性. 可以用一条通过最小二乘拟合得到的曲线来描述这一相关程度.



(a) SDMSE vs. MOS



(b) SDMAD vs. MOS



(c) SDDSSIM vs. MOS

图 4　基于 MSE/MAD/DSSIM 的空间域视觉显著性变化客观质量评估方法与 MOS 的散点图 (其中, 曲线是最小二乘拟合曲线, 每一点上的竖杠表示 95 % 置信区间.)

Fig.4　Scatter plots of MSE/MAD/DSSIM related saliency deviation based video quality metrics vs. MOS (The curve is fitted by least squares fitting and the vertical bar indicates the 95 % confidence interval.)

3.2　基于受损视频时间域的视觉显著性变化

终端用户观看的视频是图像的时间序列, 然而, 视频带来的视觉效果却不是图像的简单累加, 而是在时间域上表现出其特有的效果, 例如一些异常事件的突然发生 (爆炸、闪烁、以及图像的明显错误等) 会给观看者带来很大的视觉冲击, 但这些信息在单个图像上并不能引起人们的注意. 因此, 时间上的动态变化是视频序列区别于静止图像的重要特征. 目前的视频质量评估方法中, 考虑视频时间域失真度量的方法主要是采用直接对空间的失真度量信息在时间上求和平均[11] 或者 Minkowski 求和[22]. Itti 等[23] 分析了空间和时间域视觉注意特征的重要性, 并指出时间上的变化比其他一些静态特征更能够影响并预测人类视觉注意的变化路线. 文献 [24] 的结论指出视频时间域的质量变化频率能够影响视频的整体质量. 由于丢包引起的错误不仅在空间上扩散, 也会在时间序列上造成某些运动目标的改变或可见的异常错误, 而人眼容易被局部的异常错误所吸引, 从而改变原本期望的注意视线, 因此, 这些改变在时间域上的变化幅度能够反映视频的感知质量. 本文提出通过衡量视觉注意力在时间上的变化幅度来评估视频时间域的感知质量. 首先, 定义 $m(t)$ 为显著图的平均计算, 则 $m_s(t)$ 为第 t 帧受损帧显著图的平均计算:

$$m_s(t) = E_{(x,y)}(S(x,y,t)) \tag{6}$$

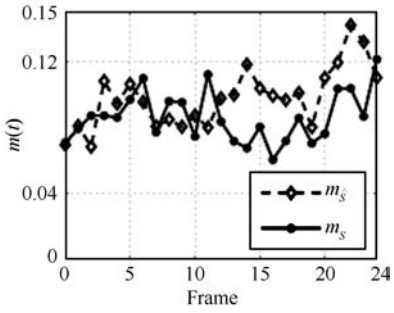
同理 $m_{\hat{S}}(t)$ 为第 t 帧相应原参考帧显著图的平均计算. 图 5 给出了 “Aircraft” 和 “Leaf” 的 $m(t)$ 图. 可以看到, 对于质量较差的 “Aircraft” (MOS = 2.45), $m_{\hat{S}}(t)$ 和 $m_S(t)$ 随着时间变化都比较剧烈; 而对于质量较好的 “Leaf” (MOS = 5), $m_{\hat{S}}(t)$ 和 $m_S(t)$ 曲线都很平缓. 因此, 通过计算 $m_s(t)$ 的标准差来度量受损视频在时间域的显著变化 (Saliency temporal variation, STV):

$$STV_s = STV_t(m_s(t)) \tag{7}$$

其中, $STV_t(\cdot)$ 是标准差计算子. 同理可以得到相应参考帧的 $STV_{\hat{s}}$ 为 $m_{\hat{S}}(t)$ 的标准差.

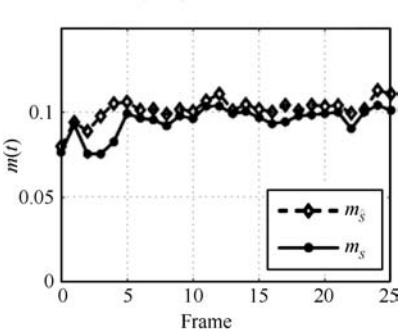
图 6 给出了 STV_s 和主观评测分数 MOS 的散点图, 可以看出二者具有很好的线性负相关度, 因此, STV_s 能够在一定程度上反映受损视频的质量. 然而, 对比图 5 (a) “Aircraft” 和图 5 (b) “Leaf” 中的 STV_s 和 $STV_{\hat{s}}$, 我们发现, 它们的变化趋势非常相近. 同样在所有测试数据集中 (见图 7), 一些质量较好的视频, 如 “Liberty” (MOS = 4.36), “Ship” (MOS = 4.00), 其 STV_s 的变化趋势和 $STV_{\hat{s}}$ 非常相近; 而一些质量较差的视频, STV_s 和 $STV_{\hat{s}}$ 的值则都一样较高, 如 “Mobile” (MOS = 2.63); 甚至一些视频的 STV_s 比 $STV_{\hat{s}}$ 的变化还大, 如 “Bus” (MOS = 2.8), “Stockholm” (MOS = 3), “Boat”

(MOS = 3.48). 因此, 如果单独用 STV_s 来度量视频的质量, 会将原参考视频的质量判断为与受损视频的质量一样差. 由此可以看出 STV_s 能够作为解释视频质量时间域动态变化的一个因素, 但却缺乏对视频空间失真信息的敏感性.



(a) 示例视频 “Aircraft”

(a) Example video “Aircraft” (MOS = 2.45)



(b) 示例视频 “Leaf”

(b) Example video “Leaf” (MOS = 5)

图 5 示例视频的 $m_S(t)$ 和 $m_{\hat{S}}(t)$

Fig. 5 $m_S(t)$ and $m_{\hat{S}}(t)$ of the example videos

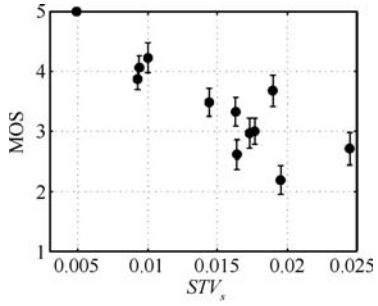


图 6 STV_s 与 MOS 的散点图

Fig. 6 Scatter plots of STV_s vs. MOS

3.3 基于空间–时间显著注意变化的评估方法

在视频处理中, 将时间与空间信息相结合的方法很多. Itti 等[23] 将空间和时间的显著信息通过线性相加进行融合, 并提出了视频的新异 (显著) 事件检测方法. 文献 [25] 中利用时间变化对空间检测的失真进行滤波, 进而提出基于时间变化的质量评估方法. 通过上一节的分析, 基于受损视频时间域的方法 STV_s 不能单独作为质量评估的度量, 因此, 在其与基于空间的注意力检测方法融合时不能简单的相加, 而需要通过一种非线性的方法实现. 另一方面, 通过实验发现, 图 8 以基于 SSIM 的评估方法为例 (评估方法都已规则化到 [0, 1]), 将 STV_s 与空间方法相乘和简单相加都可以实现一定相辅相成的效果, 并使结果优于单个评估方法. 然而, 对于一些空间与时间上注意力变化差异较大的视频序列, 如 “Whale”、“Mobile”, 将空间和时间相乘的融合方法会优于线性相加. 由此, 我们最终提出一组基于时间–空间显著注意变化, 简称为视觉注意力变化 (Saliency variation, SV) 的评估方法, 表示如式 (8):

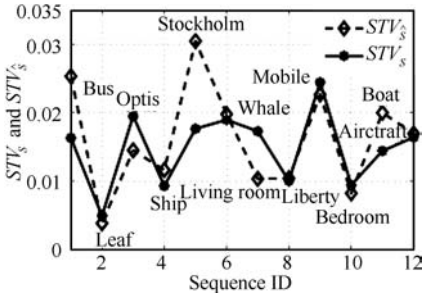


图 7 所有测试视频的 STV_s 与 $STV_{\hat{s}}$ 的对比图

Fig. 7 Comparison of STV_s and $STV_{\hat{s}}$ of all test videos

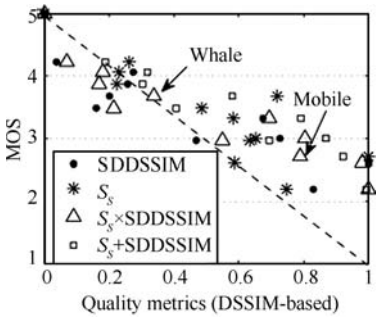


图 8 视频时间–空间信息融合方法对比 (以基于 DSSIM 的评估方法为例, 图中虚线代表基准线.)

Fig. 8 Comparison of video spatial-temporal combination methods (Taking DSSIM-based method for example, the dash line is the baseline.)

$$\begin{aligned} \text{SVMSE} &= STV_s \times \text{SDMSE} \\ \text{SVMAD} &= STV_s \times \text{SDMAD} \\ \text{SVSSIM} &= STV_s \times \text{SDDSSIM} \end{aligned} \tag{8}$$

此方法中, STV_s 不局限于与注意力变化的空间信息融合, 它还可与其他基于图像空间信息的评估方法结合.

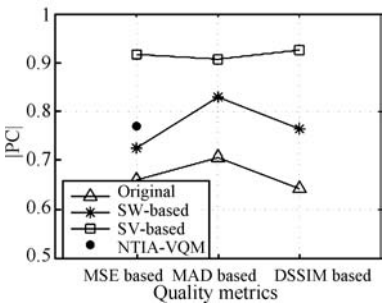
4　评估方法的性能评价及比较

为了评估本文提出的算法, 我们利用 I-SVAM 实现了目前主流的基于显著区域对错误像素进行加权的方法. 对于受损视频的错误度量同样采用 MSE, MAD 和 SSIM 三种计算, 由此定义一组基于显著性加权 (Saliency weighted, SW) 的度量方法为: SWMSE, SWMAD, SWDSSIM. 实验在第 2 节描述的 17 个视频数据上进行测试 (其中, 5 个训练视频不作为正式数据). 表 2 给出了传统没有考虑视觉信息以及没有利用视觉信息加权的方法 (Original)、基于对视觉显著错误加权的方法 (SW-based)、本文提出的基于视觉注意力变化 (SV-based) 的三种评估方法与 MOS 的 Pearson 相关系数 (Pearson correlation, PC) 以及 Spearman 等级相关系数 (Spearman′s rank order correlation coefficient, SROCC), 其中, PC 和 SROCC 的绝对值越接近 1 表明方法与主观评估结果的线性 (PC) 和非线性 (SROCC) 相关程度越高.

图 9 更直观地给出了这三组方法分别在 PC 和 SROCC 上的相关性对比. 从表 2 和图 9 中可以看出, 基于视觉显著特性的评估方法比没有考虑显著特性的评估方法在与主观评估的相关度上有很大提高. 其中, 除 SWMSE 的非线性相关度外, 所有基于 SW 的评估方法比原评估方法与主观评估的相关度更高. 而本文提出的基于视觉显著注意变化的评估方法在线性和非线性相关度上都较前两种方法有明显的提高, 其中线性相关系数都达到了 0.9 以上.

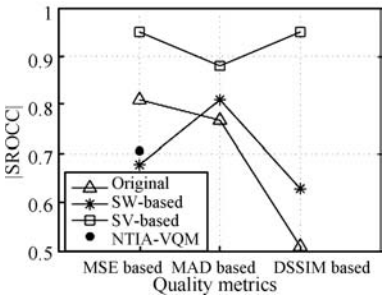
另外, 对每个视频场景的分析发现 (见图 10), 基于 SW 的评估方法能够增强发生在显著区域失真的视觉显著性、能够掩盖出现在非视觉注意区域失真的可见性, 这类方法尤其适合于评估在显著区域发生了视觉上明显失真的情况, 例如 “Optis”. 基于 SD 的评估方法通过衡量失真引起的视觉注意与期望的视觉注意之间的相对变化反映失真视频的感知质量, 这种方法不受视频本身场景的影响, 因此, 能够比基于 SW 的方法更有效地评估丢包失真视频的感知质量. 这类方法尤其适合于在背景或非视觉显著区域出现明显失真的情况, 如 “Aircraft”. 但由于基于 SW 的方法没有考虑视频的时间因素, 所以这类方法在整体表现上较差.

为了证明在视频质量评估中引入视频时间域显著注意变化度量的有效性, 我们同样将其与传统的基于空间的 MSE, MAD, SSIM 方法以及基于 SW 的方法进行相乘融合. 对比表 2 中第 2、第 3 列和第 5、第 6 列, 在 PC 和 SROCC 系数上, 所有考虑了时间注意力变化的方法都比对应的单独基于空间的质量度量方法有很大提高. 因此, 基于视频时间域显著注意变化的评估方法通过和其他基于视频空间信息的评估方法融合, 能够很好地反映受损视频时间域的信息. 但总体上看, 只有基于空间和时间注意力变化度量的结合能够使视觉显著性的空间与时间信息相辅相成, 达到最好的融合效果.



(a) 线性相关性 PC

(a) Linear correlation PC



(b) 非线性相关性 SROCC

(b) Non-linear correlation SROCC

图 9　各评估方法与主观评测结果的相关性比较

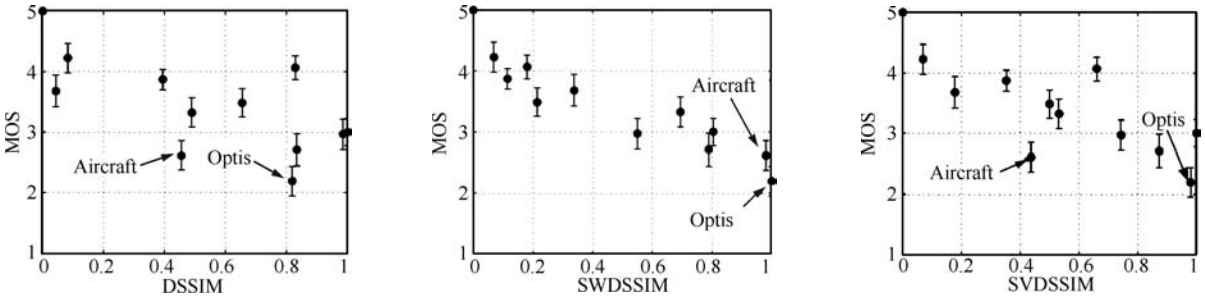
Fig. 9　Correlation performance of types of quality metrics vs. MOS

VQM (Video quality model)[26] 是由美国国家电信和信息管理局 (NTIA) 提出的视频质量评估算法 (NTIA-VQM), 并已被美国国家标准局 (ANSI) 作为国家标准采用. 在包含丢包损伤的 LIVE 数据库中, NTIA-VQM 表现出比其他视频质量评估方法更好的性能[11]. 本文同样在文中数据集中应用 NTIA-VQM, 并得到了 VQM 与 MOS 的线性与非线性相关系数 (如表 2 和图 9). 可以看到 VQM 比原 MSE/MAD/SSIM, 以及一些基于 SW 的方法 (SWMSE, SWDSSIM) 有一定提高, 但明显稍逊于所有基于 SV 的方法. 由此可以得到, 视觉显著信息在视频质量评估中起着重要的作用, 而基于视觉注意力变化的度量方法对于受丢包损伤的视频是一种有效的质量评估方法.

表 2 各评估方法与 MOS 的相关性对比

Table 2 Correlation performance of all quality metrics

MOS vs.	MSE	SWMSE	SVMSE	$STV_s\times$ MSE	$STV_s\times$ SWMSE	VQM
PC	−0.6610	−0.7252	−0.9172	−0.6632	−0.7613	−0.7701
SROCC	−0.8112	−0.6783	−0.9510	−0.8392	−0.8322	−0.7063
MOS vs.	MAD	SWMAD	SVMAD	$STV_s\times$ MAD	$STV_s\times$ SWMAD	-
PC	−0.7068	−0.8298	−0.9076	−0.7535	−0.8566	-
SROCC	−0.7692	−0.8112	−0.8811	−0.8182	−0.8531	-
MOS vs.	DSSIM	SWDSSIM	SVDSSIM	$STV_s\times$ DSSIM	$STV_s\times$ SWDSSIM	-
PC	−0.6429	−0.7645	−0.9264	−0.8404	−0.8548	-
SROCC	−0.5105	−0.6294	−0.9510	−0.7902	−0.8881	-



(a) DSSIM 与 MOS 的散点图 (b) SWDSSIM 与 MOS 的散点图 (c) SVDSSIM 与 MOS 的散点图

(a) Scatter plot of DSSIM vs. MOS (b) Scatter plot of SWDSSIM vs. MOS (c) Scatter plot of SVDSSIM vs. MOS

图 10 各方法与 MOS 的散点图 (其中竖杠表示 95 % 置信区间)

Fig. 10 Scatter plots of quality metrics vs. MOS (The vertical bar indicates the 95 % confidence interval.)

5 结论

本文针对受丢包损伤的视频序列提出了一种基于视觉注意力时间–空间变化的全参考客观质量评估方法. 其中视觉显著注意信息的检测利用 Itti 的显著区域检测模型, 并在模型中增加了基于 HR 运动检测的多尺度显著运动特征检测. 实验数据集通过对标准视频进行 H.264 编码并模拟网络单个丢包情景得到, 之后按照 ITU-R BT500-11 推荐标准实施了一个严格的主观评测实验. 通过与传统没有考虑视觉信息的评估方法, 以及目前主流的基于对视觉显著错误加权的评估方法对比, 实验结果表明, 基于视觉注意力空间–时间变化的方法与主观评估结果有很好的相关度, 对于网络丢包损伤的视频是一种有效的评估方法. 由于本文目前仅考查了单个丢包视频数据的情况, 对于存在多种丢包情况的长视频数据的质量评估, 将进一步考虑丢包的位置、长度以及宽恕效应等因素.

References

1 Wang Z, Bovik A C, Sheikh H R, Simoncelli E P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004, **13**(4): 600−612

2 Lu Liu-Ming, Lu Xiao-Yuan. Quality evaluation of video over a packet network based on packet loss. *Journal of Image and Graphics*, 2009, **14**(1): 52−58
(卢刘明, 陆肖元. 基于网络丢包的网络视频质量评估. 中国图象图形学报, 2009, **14**(1): 52−58)

3 Bouazizi I. Estimation of packet loss effects on video quality. In: Proceedings of the 1st International Symposium on Control, Communications and Signal Processing. Hammamet, Tunisia: IEEE, 2004. 91−94

4 Wolf S, Pinson M. Video Quality Measurement Techniques, NTIA-Report 02-392, National Telecommunications and Information Administration, USA, 2002

5 Babu R V, Bopardikar A S, Perkis A, Hillestad O I. No-reference metrics for video streaming applications. In: the 14th International Packet Video Workshop. California, USA, 2004

6 RUI Hua-Xia LI Chong-Rong QIU Sheng-Ke. Evaluation of packet loss impairment on streaming video. *Journal of Zhejiang University — Science A*, 2006, **7**(Suppl.1): 131−136

7 Kanumuri S, Subramanian S G, Cosman P C, Reibman A R. Predicting H.264 packet loss visibility using a generalized linear model. In: Proceedings of the IEEE International Conference on Image Processing. Atlanta, USA: IEEE, 2006. 2245−2248

8 Kanumuri S, Cosman P C, Reibman A R, Vaishampayan V A. Modeling packet-loss visibility in MPEG-2 video. *IEEE Transactions on Multimedia*, 2006, **8**(2): 341−355

9 Liu T, Wang Y, Boyce J M, Yang H, Wu Z Y. A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts. *IEEE Journal of Selected Topics in Signal Processing*, 2009, **3**(2): 280−293

10 Moorthy A K, Seshadrinathan K, Soundararajan R, Bovik A C, Cormack L K. LIVE video quality database [Online], available: http://live.ece.utexas.edu/ research/quality/live_video.html, September 17, 2010

11 Moorthy A K, Seshadrinathan K, Soundararajan R, Bovik A C. Wireless video quality assessment: a study of subjective scores and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 2010, **20**(4): 587−599

12 You J Y, Korhonen J, Perkis A. Spatial and temporal pooling of image quality metrics for perceptual video quality assessment on packet loss streams. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing. Texas, USA: IEEE, 2010. 1002−1005

13 Ninassi A, Le Meur O, Le Callet P, Barbba D. Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In: Proceedings of the IEEE International Conference on Image Processing. Texas, USA: IEEE, 2007. 169−172

14 Ling Yun, Xia Jun, Tu Yan, Yin Han-Chun. Detection of region of interest and its application in video image quality assessment. *Journal of Southeast University (Natural Science)*, 2009, **39**(4): 754−757
(凌云, 夏军, 屠彦, 尹涵春. 视觉感兴趣区的提取及其在视频图像质量评估中的应用. 东南大学学报 (自然科学版), 2009, **39**(4): 754−757)

15 Moorthy A K, Bovik A C. Visual importance pooling for image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 2009, **3**(2): 193−201

16 Radiocommunication Section of ITU, ITU-R Rec. BT500-11: Methodology for the subjective assessment of the quality of television pictures, 2002

17 Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**(11): 1254−1259

18 Bernhardt-Walther D. Saliency toolbox [Online], available: http://www.saliencytoolbox.net/, January 2, 2010

19 Borst A. Models of motion detection. *Nature Neuroscience*, 2000, **3**: 1168−1168

20 Fraunhofer Heinrich Hertz Institute, JM software 10.0 [Online], available: http://iphome.hhi.de/suehring/ tml/download/, September 28, 2011

21 Liu Y, Bu J J, Chen C, Mo L J, He K W. Multiframe error concealment for whole-frame loss in H.264/AVC. In: Proceedings of the IEEE International Conference on Image Processing. San Antonio, USA: IEEE, 2007. 281−284

22 Winkler S. *Digital Video Quality: Vision Models and Metrics*. Chichester: John Wiley and Sons, 2005. 117−120

23 Itti L, Baldi P. A principle approach to detecting surprising events in video. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 631−637

24 Zink M, Kunzel O, Schmitt J, Steinmets R. Subjective impression of variations in layer encoded videos. *Lecture Notes in Computer Science*. Berlin: Springer, 2003. 137−154

25 Ninassi A, Le Meur O, Le Callet P, Barba D. Considering temporal variations of spatial visual distortions in video quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 2009, **3**(2): 253−265

26 Pinson M H, Wolf S. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, 2004, **50**(3): 312−322

冯 欣 重庆理工大学计算机科学与工程学院讲师. 2011 年获得重庆大学计算机学院博士学位. 主要研究方向为计算机视觉、图像/视频处理和网络视频质量评估. 本文通信作者.
E-mail: fx_328@hotmail.com
(**FENG Xin** Lecturer at the College of Computer Science and Engineering, Chongqing University of Technology. She received her Ph. D. degree from the College of Computer Science, Chongqing University in 2011. Her research interest covers computer vision, image/video processing, and network video quality assessment. Corresponding author of this paper.)

杨 丹 重庆大学软件工程学院教授. 主要研究方向为图像处理, 机器视觉, 人工智能和软件工程.
E-mail: dyang@cqu.edu.cn
(**YANG Dan** Professor at the School of Software Engineering, Chongqing University. His research interest covers image processing, machine vision, artificial intelligence, and software engineering.)

张 凌 重庆通信学院讲师. 2008 年获得重庆大学计算机学院硕士学位. 主要研究方向为网络视频传输, 网络安全和软件工程. E-mail: jonent@163.com
(**ZHANG Ling** Lecturer at Chongqing Communication Institute. He received his master degree from the College of Computer Science, Chongqing University in 2008. His research interest covers video communication, network security, and software engineering.)