

基于等效 MSE 的空间-时间可伸缩码流提取方法

钱大兴, 王洪玉*, 孙文珠

(大连理工大学 电子信息与电气工程学部, 辽宁 大连 116024)

摘要: 可伸缩视频编码(scalable video coding, SVC)可以有效地将视频的码流通过不同网络传输给不同的用户. 通过对一个帧组(GOP)中所有宏块的尺寸和运动情况进行分析, 提出了等效均方误差(Eq-MSE)方法, 以此计算视频不抖动的最小帧率, 再根据网络的带宽, 确定提取子流的时间和空间增强层的层数. 这种基于视频内容的子流提取方法, 有效地弥补了当前视频编码不考虑视频内容这一不足, 在一定带宽下, 使视频的解码质量达到最优. 对比于不考虑视频内容的方法, 提出的自适应子流提取方法使重建的视频峰值信噪比在一定比特率下显著改善.

关键词: SVC; Eq-MSE; 比特流提取; 时间-空间频率

中图分类号: TN919.8 **文献标志码:** A

0 引言

随着网络技术的发展, 拥有不同处理能力的终端设备对所接收到的多媒体视频的格式要求也不同. 从处理能力有限的小屏幕手机到能力强大的高清电脑, 很难有一种统一的视频格式能够完全适用于这些连接于不同网络的不同设备. 可伸缩视频编码(scalable video coding, SVC)^[1,2]作为 H. 264/AVC^[3]的扩展部分, 由 ITU-T VCEG 和 ISO-IEC MPEG 共同开发. SVC 有广泛的研究和应用领域, 如多广播、视频监控、点到点传输等. SVC 将一个视频序列编码为一个码流, 该码流可以根据不同终端设备的处理能力、用户的需求和网络情况, 通过丢弃重要性相对较弱的部分, 来实现其可伸缩性. 终端设备接收到码流的比特数越多, 解码得到的视频质量就越好, 从而实现了视频进行一次编码即可适应所有不同终端处理器的目标.

SVC 包括时间、空间、质量可伸缩性, 多种可伸缩性还可以组合在一起. 时间可伸缩性通过 hierarchical-B 的方法来实现^[4]. 空间可伸缩性是

通过增强层结合其参考层达到更高的分辨率来实现的, 通常使用的有 pyramid 方法. 本文只讨论时间与空间可伸缩性, 质量可伸缩性将不做讨论.

不同的视频, 在一定的带宽下, 最优子流的时间和空间增强层的组合也是不同的. 如果视频的运动向量较大, 则需要较高的帧率才能使其看起来没有抖动, 所以要尽量满足时间增强层的提取; 如果视频具有丰富的纹理, 则需要使视频看起来更加清晰, 要尽量满足空间增强层的提取. 所以一种好的子流提取方法, 一定要结合视频的内容. 在一定带宽下, 多个子流所组成的集合中, 一定有且只有一个解码质量是最好的. 本文提出一种基于时间-空间可伸缩特性的子流提取方法“等效 MSE”, 自适应地选择提取最优的子流.

1 理论基础

SVC 的视频比特流由一个基本层和一个或多个增强层组成, 每个增强层都能增加时间/空间分辨率或者提高视频的质量. 图 1 说明了 SVC 基本层和增强层的关系. 最小的图片只使用了包含

基本层的子流来解码,其视频帧格式为 QCIF (176×144),帧率为 7.5 f/s;中间的图片使用了包含基本层和第一增强层的子流来解码,其视频帧格式为 CIF (352×288),帧率为 15 f/s;最大的图片的视频帧格式为 4CIF (704×576),帧率为 30 f/s,是对全码流进行解码得到的,全码流是包含一个基本层和两个增强层的码流。

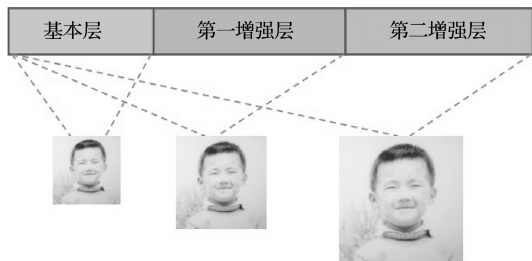


图 1 SVC 基本层和增强层

Fig. 1 SVC base layer and enhancement layers

1.1 空间频率及其作用

空间分辨率,指的是一幅图像的像素数.时间频率指的是刷新率,即单位时间内,所显示的帧数.空间分辨率组成了一个二维平面,再加上时间频率就构成了一个三维的空间.时间轴与二维平面并不是正交的,下面,本文将介绍它们之间的关系.

首先,引出空间频率的概念:

图 2 说明了图像中物体的空间频率.例如: $f_x = 3$,意味着每行有 3 个周期, $f_y = 5$,意味着每列有 5 个周期.这里采用简单的二维正弦信号是为了更清楚地说明空间频率的概念.

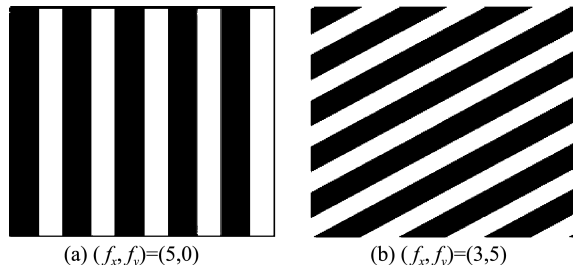


图 2 二维正弦信号

Fig. 2 2D sinusoidal signals

其次,再来看一下物体运动的成像公式:

一个线性运动的物体能够引起时间频率的变化^[5].如图 3 所示,物体在 $t = 0$ 时刻的坐标为

(x, y) ,经过时间 t ,移动距离为 $(v_x t, v_y t)$,位置为 $(x + v_x t, y + v_y t)$.或者,将物体在 t 时刻的位置记为 (x, y) ,对应于 0 时刻的位置为 $(x - v_x t, y - v_y t)$.

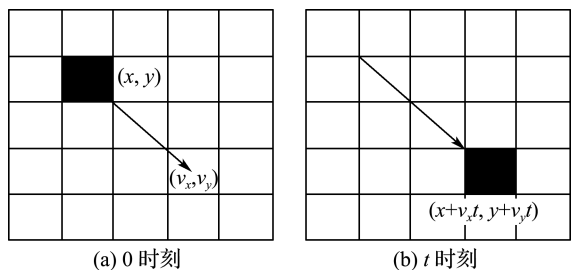


图 3 物体线性运动情况

Fig. 3 Linear movement of object

$t = 0$ 时物体的成像图形用 $\phi_0(x, y)$ 表示,它在水平和垂直方向上的速度用 v_x 和 v_y 表示,则在 t 时刻的成像图形将是

$$\phi(x, y, t) = \phi_0(x - v_x t, y - v_y t) \quad (1)$$

最后,介绍视频序列的图像与时间轴之间的关系:

对上述信号进行连续空间傅里叶变换 (CSFT),能够得到

$$\begin{aligned} \Psi(f_x, f_y, f_t) &= \iiint \phi(x, y, t) \exp(-j2\pi(f_x x + f_y y + f_t t)) dx dy dt = \\ &= \iiint [\phi_0(x - v_x t, y - v_y t) \times \\ &\quad \exp(-j2\pi(f_x(x - v_x t) + f_y(y - v_y t)))] dx dy \times \\ &\quad \int \exp(-j2\pi(f_t + f_x v_x + f_y v_y)t) dt = \\ &= \Psi_0(f_x, f_y) \int \exp(-j2\pi(f_t + f_x v_x + f_y v_y)t) dt = \\ &= \Psi_0(f_x, f_y) \delta(f_t + f_x v_x + f_y v_y) \end{aligned} \quad (2)$$

$\Psi_0(f_x, f_y)$ 是 $\phi_0(x, y)$ 的二维 CSFT. 只有在 $f_t = -f_x v_x - f_y v_y$ 时以上公式才有不为 0 的值. 这意味着如果物体以速度 (v_x, v_y) 运动,那么由物体的 (f_x, f_y) 表征的空间图案将导致一个时间频率^[5]:

$$f_t = -f_x v_x - f_y v_y \quad (3)$$

对于视频信号,时间频率依赖于 2D 图像中的位置.对于 2D 图像中固定的位置 (x, y) ,定义它的时间频率为每秒的周期数,单位为赫兹(Hz).

通过上述分析可以发现,时间频率不仅是由运动决定的,而且还取决于物体的空间频率.

1.2 空间频率与视觉感觉的关系

下面,本文将介绍在一般固定的空间频率下,所获得的最大、最小时间频率和物体运动速度方向的关系:

$$\begin{cases} f_t = -f_x v_x - f_y v_y \\ v_x = v \cdot \cos \theta \\ v_y = v \cdot \sin \theta \end{cases} \quad (4)$$

用 v 表示物体运动的速度矢量, θ 表示速度矢量与水平方向的夹角,则可以得到 $|v| = \sqrt{v_x^2 + v_y^2}$.将 v_x 和 v_y 代入 f_t ,得到 $f_t = -f_x v \cdot \cos \theta - f_y v \cdot \sin \theta$,两边对 θ 求导,得到 $f'_t = f_x v \cdot \sin \theta - f_y v \cdot \cos \theta$,令 $f'_t = 0$,可以得到:

$$\theta = \arctan(f_x/f_y) \quad (5)$$

在物体运动速度恒定的情况下,当物体运动速度的方向与空间频率的方向一致时,可以获得最大的时间频率;当物体运动速度的方向与空间频率的方向正交时,时间频率为 0.

2 算法描述

视频序列中的帧,是用分块的形式来表示的. SVC 继承了 H. 264 的分块方法,块的尺寸为 $16 \times 16, 16 \times 8, 8 \times 16, 8 \times 8$. 如果为 8×8 ,还可以继续划分为 $8 \times 4, 4 \times 8, 4 \times 4$ 三种子宏块. 利用第 1 章中提出的方法得到时间频率,需要先确定物体在水平、垂直方向上的空间频率以及在这两个正交方向上的运动速度分量. 下面介绍影响图像峰值信噪比(PSNR)的原因和非正弦波的空间频率的确定方法,此方法可以扩展成为确定一般物体空间频率的有效手段.

2.1 相同的 f_t 与不同的均方误差形成的原因

以空间频率 $(f_x, f_y) = (1, 0)$ 和 $(f_x, f_y) = (2, 0)$ 这两种特殊的情况为例,如图 4(a)、(b)所示,设图像的尺寸为 $W \times H$,两帧图像都以速度 $(v_x, v_y) = (v, 0)$ 运动,经过足够短的时间 Δt 之

后,两帧图像对应位置的像素分别为 $p_t(x, y)$ 和 $p_{t+\Delta t}(x, y)$.

利用两帧图像均方误差的计算公式:

$$MSE(x, y) = \frac{1}{WH} \sum_{x=1}^W \sum_{y=1}^H [p_t(x, y) - p_{t+\Delta t}(x, y)]^2 \quad (6)$$

图 4(a)在 t 和 $t + \Delta t$ 时刻的均方误差:

$$MSE_{4(a)}(x, y) = \frac{2}{WH} (H \times v \times \Delta t \times E)^2 = \frac{2H}{W} (v \times \Delta t \times E)^2$$

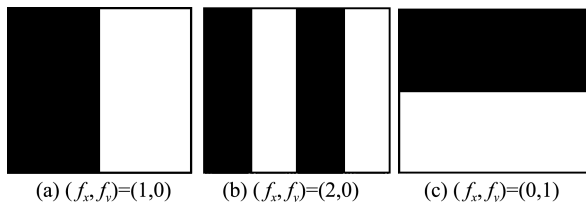


图 4 不同空间频率的二维正弦信号

Fig. 4 2D sinusoidal signals under different spatial frequencies

图 4(b)在 t 和 $t + \Delta t$ 时刻的均方误差:

$$MSE_{4(b)}(x, y) = \frac{4}{WH} (H \times v \times \Delta t \times E)^2 = \frac{4H}{W} (v \times \Delta t \times E)^2$$

其中 E 是图中黑色与白色之间的灰度差值.

图 4(b)的均方误差是图 4(a)的 2 倍,推广到更一般的情况对于空间频率 $(f_x, f_y) = (n, 0)$,在同样的运动速度和足够短的时间间隔下,它的均方误差将是 $(f_x, f_y) = (1, 0)$ 的图像的 n 倍. 同理可证,对于 $(f_x, f_y) = (0, n)$,在以上相同的条件下,均方误差是 $(f_x, f_y) = (0, 1)$ 的 n 倍.

再来看反映图像编码性能的参数峰值信噪比的计算公式:

$$PSNR_{dB} = 10 \lg(2^n - 1)^2 / MSE \quad (7)$$

$(2^n - 1)^2$ 为图像中最大可能的信号值平方, n 为每个像素的比特数. 均方误差越大,峰值信噪比就越小.

显而易见,对于引起同样时间频率 f_t 的两组图像,每组的均方误差和峰值信噪比是不相同的. 如果使用均方误差或者峰值信噪比作为自适应提取帧率的参数,是不完全合理的.

对于图 4(a)、(c), (f_x, f_y) 分别为 $(1, 0)$ 和 $(0, 1)$ 这两种情况, 如果运动速度分别为 $(v, 0)$ 和 $(0, v)$, 根据式(3), 虽然它们引起的的时间频率都是 $f_t = -v$, 但它们的均方误差显然不一致, 比值为 H/W .

2.2 等效均方误差方法

使用式(3)可以计算规则形状物体的空间频率, 进一步计算由于它的运动所引起的时间频率.

式(8)表示当前物体的运动所引起的帧率:

$$f_{tb} = -\frac{MSEf(h/H, 0)}{MSEf(1, 0)}v_x - \frac{MSEf(0, w/W)}{MSEf(0, 1)}v_y \quad (8)$$

如图 5 所示, 黑色的柱状物体所属宏块的尺寸为 $w \times h$, $MSEf(f_x, 0)$ 表示空间频率为 $(f_x, 0)$ 的图像, 由于在水平方向运动足够短的时间 Δt 之后, 与当前图像之间的均方误差的值. 类似定义 $MSEf(0, f_y)$, 表示空间频率为 $(0, f_y)$ 的图像, 由于在垂直方向运动足够短的时间 Δt 之后, 与当前图像之间的均方误差. 为了保持与式(3)一致的形式, 当前一般物体的空间频率为

$$(f_x, f_y) = \left(\frac{MSEf(\frac{h}{H}, 0)}{MSEf(1, 0)}, \frac{MSEf(0, \frac{w}{W})}{MSEf(0, 1)} \right) = \left(\frac{h}{H}, \frac{w}{W} \right) \quad (9)$$

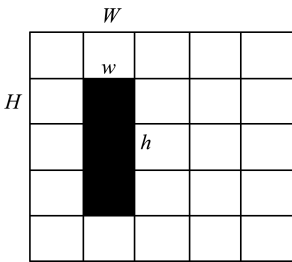


图 5 尺寸为 $W \times H$ 的图像中尺寸为 $w \times h$ 的物体

Fig. 5 $w \times h$ object in $W \times H$ picture

图 5 中的物体, 运动速度为 $(v, 0)$, 那么它与速度相同、空间频率为 $(f_x, f_y) = (1, 0)$ 的物体的均方误差的比值为 h/H , 就认为它的空间频率 $(f_x, f_y) = (h/H, 0)$. 同理, 如果运动速度为 $(0, v)$, 那么它的空间频率就为 $(f_x, f_y) = (0, w/W)$. 两帧图像中物体沿着任意方向运动所引起的时间频率为

$$f_t = \sum f_{tb} = \sum \left(-\frac{MSEf(\frac{h}{H}, 0)}{MSEf(1, 0)}v_x - \frac{MSEf(0, \frac{w}{W})}{MSEf(0, 1)}v_y \right) = \sum \left(-\frac{h}{H} \cdot v_x - \frac{w}{W} \cdot v_y \right) \quad (10)$$

其中 $\sum$ 表示取图像中所有的宏块, v_x 和 v_y 是相应宏块在水平和垂直方向上运动的速度. 如果得到一帧图像中的宏块的大小和数目, 再在同一个帧组中选取另一帧图像来计算每个宏块相应的运动向量. 以宏块为单位的运动向量与两帧之间所用时间的比值就是速度 v_x 和 v_y . 例如, 视频序列 Mobile 的 $\sum \frac{h}{H}v_x$ 和 $\sum \frac{w}{W}v_y$ 分别为 12.2 和 9.4, 它的最小不抖动帧率 (MinFR) 是 $12.2 + 9.4 = 21.6$. 对于实信号来说, CSFT 是对称的, 对于每一个 (f_x, f_y) 的频率分量, 都有一个 $(-f_x, -f_y)$ 的分量与之对应. 所以, 相应的时间频率也可以表示为 $f_x v_x + f_y v_y$. 式(10)表示了 MinFR, 它是使视频看起来运动连续不抖动的最小帧率.

根据式(3)和第 2 章的方法, 可以确定一般宏块的 f_x 和 f_y . 再根据宏块的水平和垂直方向的运动速度, 就可以确定视频序列由于空间块运动引起的时间频率, 以此来得出使运动物体看起来不抖动的最小帧率. 确定最小帧率之后, 再根据传输的比特率进一步得到适合于视频提取的空间频率. 避免出现单帧效果清晰, 但运动物体出现跳跃; 或者物体运动连续, 但单帧效果很差的情况发生, 使视频序列在一定的带宽下, 时间和空间可伸缩达到一个最佳的折中.

3 算法的仿真实验与结果分析

本文使用 Akiyo、Bridge-far、Football 和 Mobile 这 4 个视频序列来进行实验, 实验平台使用 Joint Scalable Video Model JSVM_9_12_2, 配置文件的使用方法参考 JSVM Software Manual, Version JSVM 9. 12. 2 (CVStag: JSVM_9_12_2), 帧组尺寸为 16, 编码两个序列的前两个帧组.

对于这 4 个测试序列来说, 由于 Football 和 Mobile 运动较为剧烈, 两帧之间物体运动变化较

大,对帧率的要求比较高;而 Akiyo 和 Bridge-far 运动较为缓和,物体运动变化较小,可以适当减小帧率,将节省下来的带宽分配给空间分辨率.

使用 (D,T) 表示层提取的情况, D 表示空间可伸缩的阶数, $D=0$ 时分辨率为 QCIF, $D=1$ 时分辨率为 CIF; T 表示时间可伸缩的阶数,所对应的帧率见表 1.

表 1 时间可伸缩的阶数对应的帧率
Tab.1 Frame ratio corresponding to temporal scalability stages

T	帧率/($f \cdot s^{-1}$)
0	1.875
1	3.750
2	7.500
3	15.000
4	30.000

3.1 主观质量比较

图 6、7 是 Mobile 和 Bridge-far 在不同的空间频率和帧率下的图像.

本文采用传输的比特率比较接近的两种提取方法对编码后的视频序列进行提取.对于 $(D,T)=(0,4)$ 的提取序列,虽然空间频率看上去要差一些,但其中物体的运动是连续的,并没有跳跃的感觉;而 $(D,T)=(1,2)$ 的提取序列,虽然单帧质量

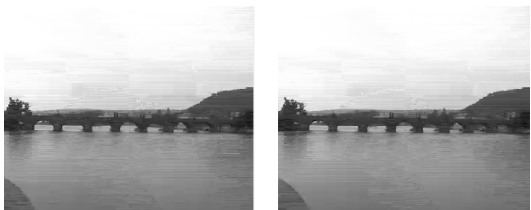


(a) $(D,T)=(1,2)$



(b) $(D,T)=(0,4)$

图 6 测试序列 Mobile 解码后序列最初的 2 帧
Fig.6 First two frames of Mobile decoded



(a) $(D,T)=(0,4)$



(b) $(D,T)=(1,3)$

图 7 测试序列 Bridge-far 解码后序列最初的 2 帧
Fig.7 First two frames of Bridge-far decoded

接近于原始序列,但其中物体的跳跃感,让更多的人主观上感到不可接受.

对于 Bridge-far 来说,帧内的物体几乎是没有什么运动的.所以,人眼更容易接受对于空间频率较高的提取方法.这两种提取方法的比特率比较接近, $(D,T)=(0,4)$ 的比特率(197.70 kbps)要稍高于 $(D,T)=(1,3)$ 的比特率(156.30 kbps).

3.2 客观质量比较

本文使用以上提到的方法,对 Akiyo、Bridge-far、Football、Mobile 这 4 个测试序列进行不同的子流提取,并进行比较.每个测试序列编码为 2 个空间层、5 个时间层,2 个空间层的 QP 值设为 36 和 32.分别提取每个测试序列不同的时空子流进行比较,所提取子流的比特率应该尽量接近.结果见表 2.

表 2 视频序列客观比较结果
Tab.2 Objective comparison of video sequences

视频序列	(D,T)	峰值信噪比/dB	比特率/kbps
Akiyo	(0,3)	12.01	284.0
	(1,1)	23.73	163.2
Bridge-far	(0,4)	33.13	197.7
	(1,3)	37.54	156.3
Football	(0,4)	31.65	845.3
	(1,3)	15.50	1 248.9
Mobile	(0,4)	28.48	1 344.8
	(1,2)	17.72	1 025.3

显然,对于不同的序列,在带宽接近的条件下,对于提取不同的子流所恢复出的视频,客观方面可以带来不同的峰值信噪比,主观方面人眼所更希望接受的视频质量也不同.

4 结 语

本文提出一种基于视频内容的 SVC 码流自适应提取方法. 首先,通过等效 MSE 方法,将空间频率的概念,从规则的 2D 正弦信号推广到一般情况,使其具有普遍性,适合于实际的应用. 其次,根据画面的运动情况,得出其引起的 MinFR. 在带宽受限的条件下,以此作为视频的最小帧率,避免视频中物体的运动产生抖动. 最后,在满足物体连续运动的条件下,尽量增加每一帧的空间频率,来提高清晰度. 实验结果表明,在带宽一定的条件下,根据不同的视频内容来提取最优的子流,使视频质量在主观和客观方面都得到了明显的提高.

参考文献:

[1] Schwarz H, Marpe D, Wiegand T. Overview of the scalable video coding extension of the H. 264/AVC

standard [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2007, 17(9):1103-1120.

[2] Sullivan G, Luthra A, Wiegand T. Text of ISO/IEC 14496-10: 2005/FDAM 3 scalable video coding, Joint Video Team (JVT) of ISO-IEC MPEG & ITU-T VCEG [C] // **13 Meeting : Palma de Mallorca**. Palma de Mallorca:[s n], 2004.

[3] CHEN Ying, ZHAI Jie-fu, YIN Peng, *et al.* ISO/IEC ITU-T Rec. H264: Advanced video coding for generic audiovisual services, Joint Video Team (JVT) of ISO-IEC MPEG & ITU-T VCEG [C] // **72nd MPEG Meeting**. Busan:[s n], 2005.

[4] Schwarz H, Marpe D, Wiegand T. Analysis of hierarchical B pictures and MCTF [C] // **Proceedings of the IEEE International Conference on Multimedia & Expo, ICME '06**. Piscataway:IEEE, 2006.

[5] WANG Y, Ostermann J, ZHANG Y Q. **Video Processing and Communications** [M]. New York: Prentice-Hall, 2001.

An extraction algorithm based on equivalent MSE method for temporal-spatial SVC

QIAN Da-xing, WANG Hong-yu*, SUN Wen-zhu

(Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: Scalable video coding (SVC) is an efficient and effective solution to stream video over heterogeneous networks to different clients. An equivalent MSE (Eq-MSE) method is proposed by analyzing the motion and texture in a group of picture (GOP) to calculate the minimum frame rate of non-shaking video. According to the content of a video sequence which is temporal or spatial, it determines how many temporal and spatial enhancement layers should be extracted. The method has revised the drawback which doesn't consider the video content effectively and improved video decoding performance at some band width. Experimental results show that the adaptive extraction method for sub-stream improves the PSNR of reconstructed video significantly under a certain bit rate.

Key words: SVC; Eq-MSE; bit stream extraction; temporal-spatial frequency