一种基于运动特征的快速有效的视频镜头全拼图方法

梅 涛[1] 周荷琴[1] 华先胜[2]

[1](中国科学技术大学多媒体计算与通信教育部-微软重点实验室，合肥 230027)

[2](微软亚洲研究院，北京 100080)

摘 要 由于全拼图能比关键帧提供更多的视觉信息，因此它已经成为视觉计算中一个重要的分析工具。为了提高全拼图的质量和拼图速度，提出了一种基于运动特征的快速有效的全拼图生成方法。该方法首先给定一个视频镜头，并基于运动相位熵的分析方法决定该镜头内容是否适合生成全拼图；然后，对于适合生成全拼图的镜头，通过构造全局运动路径的方法，仅需要挑选全部视频帧的一个子集用来生成高质量的全拼图。实验结果表明，与传统的全拼图方法相比，该新方法在提高全拼图的视觉质量的同时，显著地降低了计算时间。

关键词 视频全拼图 运动相位熵 全局运动路径 视频内容表征

中图法分类号：TP391.4 **文献标识码**：A **文章编号**：1006-8961(2007)03-0511-06

A Fast and Efficient Video Shot Mosaicing Approach Based on Motion Characteristics

MEI Tao[1], ZHOU He-qin[1], HUA Xian-sheng[2]

[1](*MOE-MS Key Laboratory of Multimedia Computing and Communication, University of Science and Technology of China, Hefei* 230027)

[2](*Microsoft Research Asia, Beijing* 100080)

Abstract Presenting more comprehensive visual information than key-frame, mosaic has been an important element for many vision tasks. This paper proposed a fast and efficient video mosaicing approach based on motion characteristics, which investigated two important issues affecting the quality of mosaic. First, given a video shot, a motion entropy based method is to determine whether the visual content of a shot is suited to be represented by a mosaic. Second, if so, a suitable subset of frames in this shot is selected for efficient and effective mosaicing by constructing a global motion path. Experiments compared to traditional mosaicing have proved that with this approach, not only the visual quality of mosaic is significantly improved, but also the computational time is remarkably reduced.

Keywords video mosaic, motion entropy, global motion path, video content representation

1 引 言

全拼图(mosaic)是一种将连续的视频帧合成为一张静态图片的技术。与传统的基于关键帧的视频镜头内容表征方式相比，全拼图由于包含了整个镜头的空间和摄像机运动信息，因此在计算机视觉领域有着广泛的应用。

如图 1 所示，给定一组连续的视频帧，如镜头(shot)，全拼图的生成一般包括全局运动估计(global motion estimation, GME)和全拼图合成(integration)两个步骤[1,2]。在全局运动估计中，通常由一些参数模型表征全局运动，如平移、仿射、透视模型等，而在全拼图合成中，则通常先取其中一帧

基金项目：多媒体计算与通信教育部-微软重点实验室开放基金资助项目(05071805)

收稿日期：2005-10-28；**改回日期**：2006-02-13

第一作者简介：梅涛(1978～)，男。2001 年 7 月于中国科技大学获工学学士学位，现为中国科学技术大学自动化系博士研究生。主要研究方向为图像和视频的内容分析，已发表学术论文 10 余篇。E-mail：meit@ustc.edu.cn

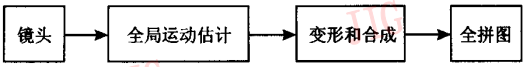


图 1 传统的视频全拼图生成方法

Fig. 1 Typical video mosaicing

作为参考平面;然后将其他帧根据运动估计的结果相对于参考帧逐帧做变形(wrapping);最后将所有变形的帧合成为一张静态图片。图 2 显示了全拼图的合成过程。

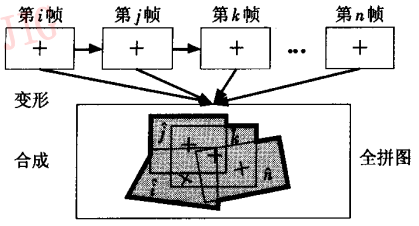


图 2 全拼图的合成

Fig. 2 Mosaic integration

尽管全拼图技术已经被应用于计算机视觉的很多方面,如视频压缩[1]、虚拟现实[2]、视频表征[3]和视频检索[4]等,但是以下两个影响全拼图质量的问题却很少被提及:

(1) 满足什么条件的视频镜头,其内容适合用一张全拼图进行表征?为简单起见,本文将满足此条件的镜头称为“m-镜头”。由于前景物体对全局运动估计的精度产生影响(如特写镜头中的前景),因此并非所有的镜头都可以生成一张没有视觉失真的全拼图。

(2) 在镜头中应该挑选哪些视频帧(本文称之为“m-帧”),使得在生成更高质量全拼图的同时,还可尽可能地减少计算时间?这里镜头被定义为摄像机的一次连续拍摄[5]。由于摄像机的连续运动,使镜头中的连续两帧之间通常存在较大重叠,因此实际上并非所有的视频帧都需要被用来生成全拼图。在传统的全拼图生成方法中,由于所有帧均参与运动估计和全拼图的合成,因而十分耗时;而且由于还需根据运动估计结果相对参考帧做逐帧变形,这样离参考帧时间上越远的帧,其运动估计的累计误差越大,变形之后的失真也就越明显,因而降低了全拼图的视觉效果。

为解决以上问题,本文提出了一种基于运动特征的快速有效的全拼图生成方法。目标是在提高全拼图质量的同时,显著减少计算时间。生成全拼图时,首先根据镜头的运动熵信息判断该镜头是否是“m-镜头”,对于“m-镜头”,则要构造一条近似的全局运动路径,并在此路径上挑选一个帧子集(即“m-帧”),用以生成高质量的全拼图。

2 方 法

图 3 给出了本文基于运动特征的快速全拼图生成方法的框图。和传统的全拼图生成算法相比(如图 1 所示),本文的方法在进行全局运动估计之前,需先对镜头内部的运动特征进行分析,即首先通过全局运动滤波器来消除异常运动的影响,然后进行“m-镜头”的检测,最后建立全局运动路径,并在此路径上选取“m-帧”。

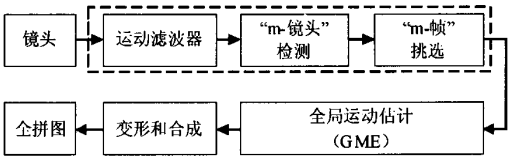


图 3 本文的全拼图生成方法

Fig. 3 Proposed video mosaicing

2.1 视频镜头的运动特征分析

具有运动特性是视频区别于其他媒体(如图像、文本等)的本质特征。通常,运动矢量场(motion vector field, MVF)被认为是光流场的一个近似。考虑到计算效率,本文假定所有的视频编码均为 MPEG 格式,MVF 可以在 MPEG 流中直接提取获得。

众所周知,在 MVF 中,运动矢量的相位表示了宏块的运动方向,而相位的分布则反映了运动矢量的空间连续性。由于相位的空间连续性可以用相位熵(phase entropy)来定量表示[6],因此对于一个 MVF,若其相位的空间连续性越好(即相位熵值越小),则越表明与该 MVF 对应的视频帧存在一致的、可测的全局运动。进一步,如果一个镜头中所有帧的相位分布均具有良好的空间连续性(数学上可描述为相位熵的平均值较低),则表明该镜头内部的全局运动较为一致,其内容就适合用全拼图的形式来进行表征,该镜头即为“m-镜头”;反之,则表明该镜头中的 MVF 较为杂乱,没有一致的全局运动,镜头内容趋向于特写镜头,就不适合生成全拼图。

对于“m-镜头”,可以先利用累加相邻两帧之间

的全局运动矢量的方法,通过构造一个全局运动路径用来近似地表征镜头内摄像机的运动;然后,在此全局运动路径上就可以挑选出适合生成全拼图的“m-帧”。

2.2　全局运动滤波器

由于运动矢量场 MVF 仅仅是光流场的近似,且矢量场中的异常采样点会导致全局运动计算过程中的误差,因此在计算全局运动之前,必须消除这些异常采样的影响。为此,本文设计了一个 3 维的时空滤波器。该滤波器的空间窗口大小为 $W_S \times W_S$(下角 S 代表 spatial),时间窗口长度为 W_t(下角 t 代表 temporal)。对于 MVF 中位置为(i, j)的宏块,经过滤波之后的运动矢量 $\hat{V}_{i,j}^{\text{motion}}$ 为该 3 维滤波器中所有宏块的运动矢量 $V_{k,l}^{\text{motion}}$ 的中值,即

$$\hat{V}_{i,j}^{\text{motion}} = \text{median}(V_{k,l}^{\text{motion}}), 1 \leqslant k, l \leqslant W_S^2 W_t \quad (1)$$

以下所有运动特征的分析都是基于 3 维时空滤波之后的 MVF 进行的。

2.3　基于运动熵的“m-镜头”检测

如 2.1 节所述,由于相位熵反映了 MVF 的空间连续性,这里对运动熵的分析仅限于运动矢量的相位熵。如果将一个运动矢量看作是一个随机变量,则一个 MVF 中的相位连续性可以用相位的熵值来衡量。该相位熵从直观上反映了摄像机运动相对于物体运动的比率,若相位熵值越大,则 MVF 空间连续性越差,即表明物体运动在 MVF 中占主导地位。这样一个镜头中的相位连续性可由所有 MVF 相位熵的均值得到。

“m-镜头”检测时,首先从一个经过 3 维滤波之后的 MVF 计算出相位直方图 H_i^{phase},而归一化的相位熵定义为

$$E^{\text{phase}} = -\frac{1}{\log n}\sum_{i=1}^{n} p_i \log(p_i)$$
$$p_i = H_i^{\text{phase}} \Big/ \sum_{j=1}^{n} H_j^{\text{phase}} \quad (2)$$

其中,$i = 1, \cdots, n$,n 为相位直方图的量化区间,$\log n$ 为归一化因子。

一个镜头的相位熵可以定义为所有帧的相位熵的均值。一个具有 M 帧的镜头被检测为“m-镜头”,当且仅当其平均相位熵小于一个预定义的阈值 τ:

$$\bar{E}^{\text{phase}} = \frac{1}{M}\sum_{i=1}^{M} E_i^{\text{phase}} < \tau \quad (3)$$

τ 的取值大小决定了一段视频序列中检测到的“m-镜头”的数量。对 τ 的取值,可以通过预定义确定,也可以通过对特定视频序列中所有镜头的平均相位熵值进行统计,再根据期望的“m-镜头”的数量来确定。

2.4　基于全局运动路径的“m-帧”挑选

在传统的全拼图生成方法中,运动估计和全拼图合成均基于连续相邻两帧进行。鉴于相邻帧之间的重叠并没有被很好的利用,由此产生了以下两个问题:(1)全拼图的视觉质量容易受到逐帧变形的累积误差的影响,特别是当用以生成全拼图的视频帧数量过多时,则累积误差越大,其生成的全拼图质量越低;(2)由于所有帧均不加考虑地参与计算,因此计算开销过大。为此,本文通过构造镜头内摄像机的全局运动路径的方法挑选出一组最佳的视频帧(“m-帧”)。其目标是尽可能地减少用以生成全拼图的视频帧的数量,以便在减少全拼图计算时间的同时提高视觉质量。

为构造近似的全局运动路径,应首先对于每一个 MVF,计算该帧摄像机运动的表征矢量(称为“r-motion”)。“r-motion”可近似地代表 MVF 的全局运动矢量,并且其幅值和相位这两个分量在 MVF 中应该具有最大的出现概率。为此,应先分别对 MVF 中所有宏块运动矢量的两个分量(相位和幅值)建立统计直方图;然后将 0 ~ 2π 范围内的相位量化成 n 个区间,而对幅值则量化成 m 个区间;最后分别在两个直方图中选择概率最大的区间所对应的相位和幅值,作为该帧“r-motion”的两个分量。

计算出镜头中所有 MVF 的“r-motion”之后,全局运动路径就可以在笛卡尔坐标系中通过将“r-motion”进行逐帧累加来得到。图 4 给出了一个基于全局运动路径挑选“m-帧”的例子。节点 O、A、B、…、H分别代表视频帧,$\overrightarrow{OA}$、$\overrightarrow{AB}$、…、$\overrightarrow{GH}$表示两帧之间的“r-motion”,则全局运动路径为 $OABCDEFGH$。挑选“m-帧”时,从第 1 帧 O 开始,先对每个“r-motion”进行矢量累加,分别得到累积运动矢量为 $\overrightarrow{OA}$、$\overrightarrow{AB}$、$\overrightarrow{OC}$、…,再将每一次累积运动矢量的幅值与一个预先定义的累积运动参量阈值 T 进行比较:如果该累积运动矢量的幅值超过 T,则停止累加过程,并将当前节点和此次累加过程的起始点都标记为“m-帧”,然后将当前节点作为下一次累加过程的起始点,开始下一次的累加;如果该累计运动矢量的幅值小于 T,则继续此次累加过程,直至该累计运动矢量的幅值超过 T。图 4 中,由于经过 4 次累加,$|\overrightarrow{OD}| > T$,所以 O 和 D 帧被选为“m-帧”;同理,由于

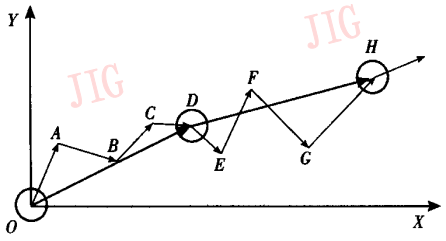


图 4 基于全局运动路径的“m-帧”挑选

Fig. 4 “m-Frame” selection on global motion path

$|\overrightarrow{DH}| > T$,所以 H 帧也被选为“m-帧”。最后,所有被挑选的“m-帧”即可用来生成全拼图。阈值 T 控制着被挑选的“m-帧”的数量。T 越大,“m-帧”越多;反之,则“m-帧”越少。

2.5 全拼图的生成

全拼图的生成包括全局运动估计和全拼图合成两个步骤。与传统全拼图生成方法使用所有帧进行运动估计和全拼图合成不同的是,本文方法仅将挑选出来的一个帧子集(即所有“m-帧”)用于全拼图的生成。

生成全拼图的关键步骤是对两帧之间的全局运动进行准确的描述。一般来说,当与摄像机平面垂直的方向没有运动或者运动很小时,全局运动可以用以下 8 参数的近似平面投影变换表示:

$$\begin{cases} v_x = a_0 + a_1 x_i + a_2 y_i + a_6 x_i x_i + a_7 x_i y_i \\ v_y = a_3 + a_4 x_i + a_5 y_i + a_6 x_i y_i + a_7 y_i y_i \end{cases} \tag{4}$$

其中,$a_i(i=0,\cdots,7)$ 表示运动模型的参数,(v_x, v_y) 表示像素点 (x_i, y_i) 的光流矢量。本文采用鲁棒的多分辨率运动估计算法[7]对式(4)中的参数进行估计,而在进行全拼图合成时,则使用时序中值滤波的方法[1]来得到镜头内容的背景全拼图。

3 实验结果

为验证本文方法的效果,利用不同的视频镜头进行了测试。实验中所有的视频编码格式均为 MPEG-1,分辨率为 352×288,帧率为 25fps。所有实验均在一台 Pentium IV3 GHz,1GB RAM 的机器上进行。根据大量视频镜头的平均相位熵和累计运动矢量的统计结果,在检测“m-镜头”时,取阈值 $\tau = 0.85$;在构造全局路径时发现,T 取值范围为 6~15时,生成的全拼图不会产生明显的视觉失真。全拼图合成时,取所有“m-帧”的中间帧为参考帧。

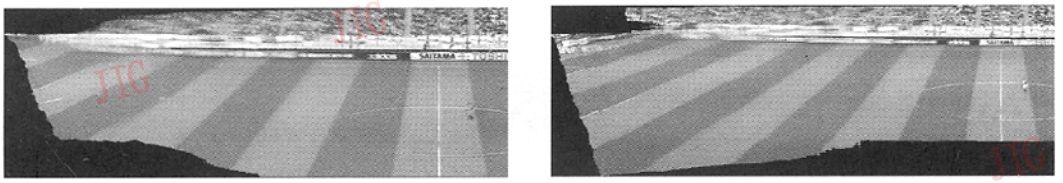
表 1 列出了本文方法和传统全拼图方法的实验结果对比。为了便于比较,在传统方法中,本文设取的帧间隔 d 分别为 $d=1$ 和 $d=3$;在本文方法中,T 分别取 10 和 15。由于镜头 1~3 的平均相位熵 $\overline{E}^{\text{phase}}$ 超过 τ,没有被检测为“m-镜头”,因此不适合生成全拼图。由此可以看出,和传统方法相比,本文的方法虽然增加了运动滤波、检测“m-镜头”和挑选“m-帧”等 3 个步骤,但是这 3 个步骤的时间开销非常小(与运动估计和全拼图合成这两步的时间开销相比可以忽略),并且由于减少了用于运动估计和全拼图合成的帧数,因而计算时间显著减少。例如,在 T 取 10 和 15 的情况下,镜头 8 的计算时间由 $d=1$ 时的2 820s,分别减少到 245s 和 384s,减少的幅度分别达到了 91% 和 86%。

由于用于生成全拼图的帧数减少,使全拼图合成时逐帧变形的累积误差也减少,因此全拼图的视觉质量有明显提高。图 5 和图 6 分别显示了镜头 4 和镜头 8 用两种方法生成的全拼图。由该两图可以看出,利用本文方法生成的全拼图视觉失真更小、质

表 1 本文方法和传统全拼图生成方法的比较

Tab. 1 Evaluation of typical and proposed mosaicing

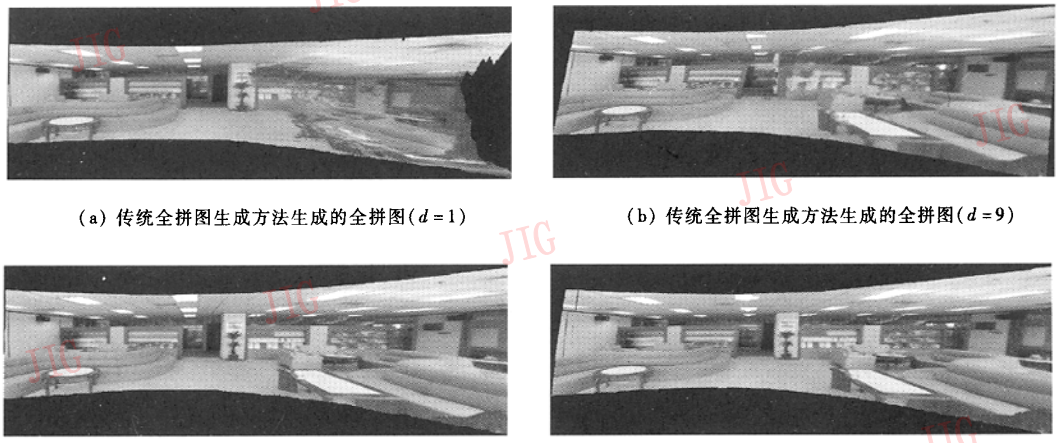
镜头	镜头帧数	传统方法				本文方法					
		$d=1$		$d=3$		$T=10$			$T=15$		
		帧数	时间(s)	帧数	时间(s)	“m-帧”数	时间(s)	$\overline{E}^{\text{phase}}$	“m-帧”数	时间(s)	$\overline{E}^{\text{phase}}$
1	47	–	–	–	–	–	–	0.92	–	–	0.92
2	50	–	–	–	–	–	–	0.91	–	–	0.91
3	77	–	–	–	–	–	–	0.92	–	–	0.92
4	347	347	3 739	114	918	53	338	0.65	35	259	0.65
5	200	200	1 698	67	671	41	355	0.81	25	290	0.81
6	192	192	3 701	64	884	40	557	0.43	26	563	0.43
7	392	392	6 281	131	695	42	237	0.75	27	226	0.75
8	349	349	2 820	117	855	38	245	0.26	28	384	0.26



(a) 传统方法($d=1$)　　(b) 本文方法($T=10$)

图 5　镜头 4 的全拼图实例

Fig. 5　Example of the mosaics of shot 4



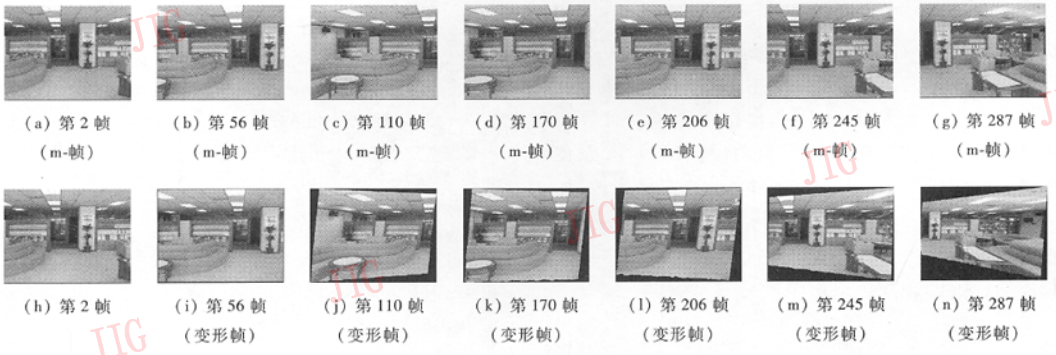
(a) 传统全拼图生成方法生成的全拼图($d=1$)　　(b) 传统全拼图生成方法生成的全拼图($d=9$)

(c) 本文方法生成的全拼图($T=10$)　　(d) 本文方法生成的全拼图($T=15$)

图 6　对镜头 8 采用传统全拼图方法和本文方法生成的全拼图的比较(摄像机实际路径:中间→左→中间→右)

Fig. 6　A comparison between typical and proposed mosaic of shot 8(The actual camera motion path: middle→left→middle→right)

量更高。从图 6 可以看出,图 6(c)和图 6(d)与图 6(a)和图 6(b)相比,失真更小,图 6(a)和图 6(b)的右半部分(对应于全局路径的末端)的失真非常明显。这是由于位于全局路径末端的视频帧在做逐帧变形时的累积误差较大,失真较明显,因而在合成时影响了最终生成的全拼图的视觉质量。图 7 给出了镜头 8 中一些“m-帧”和对应的变形帧的示例。图 8 给出了镜头 8 中第 305 帧在不同方法和参数下的变形结果。在图 8 中,用传统方法生成全拼图,时间上距离参考帧(第 170 帧)较远的帧,其变形失真较大。例如,从第 305 帧到第 170 帧,经过了 135 次变形(图8(b));相反,由于利用本文的方法($T=15$),变形



(a) 第 2 帧 (m-帧)　(b) 第 56 帧 (m-帧)　(c) 第 110 帧 (m-帧)　(d) 第 170 帧 (m-帧)　(e) 第 206 帧 (m-帧)　(f) 第 245 帧 (m-帧)　(g) 第 287 帧 (m-帧)

(h) 第 2 帧 (变形帧)　(i) 第 56 帧 (变形帧)　(j) 第 110 帧 (变形帧)　(k) 第 170 帧 (变形帧)　(l) 第 206 帧 (变形帧)　(m) 第 245 帧 (变形帧)　(n) 第 287 帧 (变形帧)

图 7　镜头 8 的一些“m-帧”和其对应的变形帧

Fig. 7　Example of m-frames and corresponding wrapped frames in shot 8

(a) 原始帧 (b) 变形帧($d=1$) (c) 变形帧($d=9$) (d) 变形帧($T=10$) (e) 变形帧($T=15$)

图 8 不同方法和参数条件下的变形帧(镜头 8 第 305 帧)比较

Fig. 8 A comparison of the wrapped frames(305th frame of shot 8) with different methods and parameters

次数减少到 16 次(图 8(e)),因此生成的全拼图视觉效果(图 6(d))要比传统方法好(图 6(a))。由图 8(c)和 8(d)可见,在 $d=9$ 和 $T=10$ 的情况下,尽管传统方法和本文方法都使用了相同数量的帧用于全拼图的生成(39 帧),但是由于传统方法所取帧间隔过大,没有考虑到两帧之间所能允许的最大运动,因此生成的全拼图(图 6(b))出现了明显的失真(与图 6(c)相比)。实验结果表明,利用本文提出的方法生成的全拼图,其视觉效果比传统方法有了较大改进。

作为本文全拼图方法的一种应用,图 9 展示了对一段家用视频利用全拼图技术进行内容表征的结果。镜头 17、24 和 30 被检测为"m-镜头",其视觉内容可用全拼图的形式进行表征。由此可见,本文的方法为视频内容的表征提供了一种有效的补充。

(a) 镜头 01 (b) 镜头 10 (c) 镜头 13 (d) 镜头 17 (m-帧) (e) 镜头 23 (f) 镜头 24 (m-帧) (g) 镜头 30 (m-帧)

图 9 利用本文的全拼图生成方法对一段家用视频内容进行表征

Fig. 9 A mosaic-based representation of a home video sequence

4 结 论

本文提出了一种新的基于运动特征的快速全拼图生成方法。该方法首先通过对运动矢量相位熵进行分析来判断一个镜头的内容是否适合用全拼图的形式进行表示;然后对满足此条件的镜头构造摄像机的全局运动路径,再在此路径上选取一个有效的帧子集用以生成全拼图。同传统全拼图生成方法进行对比的实验结果表明,由于用于进行运动估计和全拼图合成的帧数减少,因此全拼图生成的计算时间显著降低;同时,由于累积运动误差的减少,致使全拼图的视觉质量同传统方法相比,也有较大改进。今后的研究重点在于利用最短路径算法寻找当前帧相对于参考帧做变形的最短路径,以便进一步减小运动估计误差和改进全拼图质量。

参考文献(References)

1 Irani M, Anandan P, Hsu H. Mosaic based representations of video sequences and their applications [A]. In: Proceedings of IEEE International Conference on Computer Vision [C], Boston, MA, USA, 1995: 605~611.

2 Szeliski R. Video Mosaics for Visual Environments [J]. IEEE Computer Graphics and Applications, 1996, **16**(2): 22~30.

3 Hidalgo J R, Salembier P. Robust segmentation and representation of foreground key-regions in video sequences [A]. In: Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing[C], Salt Lake City, UT, USA, 2001: 1565~1568.

4 Mei T, Ma Y F, Zhou H Q, *et al*. Sports Video Mining with Mosaic [A]. In: Proceedings of IEEE International Multi-Media Modeling Conference[C], Melbourne, Australia, 2005: 107~104.

5 Zhang H J, Kankanhalli A, Smoliar S W. Automatic partitioning of full-motion video[J]. Multimedia System, 1993, **1**(1): 10~28.

6 Ma Y F, Lu L, Zhang H J, *et al*. A user attention model for video summarization[A]. In: Proceedings of ACM International Conference on Multimedia[C], Juan-les-Pins, France, 2002: 533~542.

7 Odobez J M, Bouthemy P. Robust multiresolution estimation of parametric motion models[J]. Journal of Visual Communication and Image Representation, 1995, **6**(4): 348~365.