基于字典选择的机器人在线场景语义浓缩

黄疆坪^{①②}, 丛杨^{②*}, 高宏伟^①, 唐延东^②, 于海斌^②

① 沈阳理工大学信息科学与工程学院, 沈阳 110159;

② 中国科学院沈阳自动化研究所, 机器人学国家重点实验室, 沈阳 110016

* 联系人, E-mail: congyang@sia.cn

2013-08-20 收稿, 2013-10-11 接受

国家自然科学基金(61105013, 61375014)资助

摘要 视频数据是信息存储的重要手段, 但是很多时候(如机器人运动时采集的视频、监控相机记录的视频等), 视频帧与帧之间相似度高, 造成信息冗余, 为信息的存储、查询、传输等带来困难. 本文设计了一种语义浓缩算法, 通过提取视频中的关键帧, 实现快速、准确感知视频内容的目的. 该算法首先采用背景建模方法从原始视频中提取包含重要信息的连续前景视频段, 去除信息量较少的背景帧, 然后通过采用视频语义分割方法将前景视频段分割得到一系列子视频, 将从子视频的每一帧中提取出的特征构成原始字典, 采用字典选择方法提取出关键帧. 本文方法针对导航/监控视频, 使用背景建模过滤无用信息, 通过视频语义分割方法可处理不限长度的视频, 而所选取的图像特征和字典选择方法则灵活有效地提取出有意义的关键帧. 在实验中, 通过对移动/固定平台视频的测试, 并与人工标识及其他语义浓缩方法进行比较, 证明本算法的重构误差较小, 可以准确地提取视频中的关键帧.

关键词

字典选择
语义浓缩
机器人
在线学习

自主移动机器人在室外或室内环境作业时, 其视觉系统将实时采集并记录海量的视频数据, 以用于实时导航并形成知识库, 进一步应用于目标识别等高层认知过程. 此类视频数据包含很多冗余信息和干扰信息, 这将导致以下问题: (1) 其中夹带的冗余和干扰信息将降低后期处理的准确性. (2) 海量数据的处理耗费很多 CPU 资源, 从而导致处理速度缓慢, 降低系统的实时性. (3) 海量数据的存储占用大量的存储空间. 基于上述因素, 亟待开发一种有效的、能够自动剔除视频数据冗余信息的算法机制. 近年来, 该问题也引起了国外研究机构的重视, 如 Carnegie Mellon University 的 Sofman^[1]提出一系列在线算法, 提高自主移动机器人的导航能力, 更好地利用数据和资源, 降低使用风险; McGill University 的 Gridhar 等人^[2]将浓缩图像看作真实世界的先验假设, 再用集合论计算图像的新颖度, 从而得到实时摘要;

SIA 的 Cong 等人^[3]在稀疏一致性约束下, 将个人视频摘要看作一个字典选择问题, 使得原始视频能使用具有代表性的字典获得最好的重建.

此外, 在智能视频分析中, 也有类似的问题, 监控视频中大部分都是静止的背景图像, 即使前景帧之间相似度也很高, 即存在内容冗余. 夹杂大量冗余信息的监控视频, 其传输、存储所消耗的时间和空间很大, 既降低监控系统的设备利用率, 也形成巨大的浪费. 国外学者也对此做出一系列研究, 例如, (1) 使用稀疏表示的字典选择方法^[4]从一堆相似的图片中挑选出具有代表性的图片, 但其难以处理长达数小时的监控视频. (2) 个人语义浓缩^[5]将原始视频切割成包含事件的子的视频, 依据视频中所出现目标的重要程度来提取关键帧, 但它对于场景单一的监控视频效果并不好. (3) 通过场景聚合实现的语义浓缩^[6]将视频中出现的若干相关场景拼接在一起, 损失

引用格式: 黄疆坪, 丛杨, 高宏伟, 等. 基于字典选择的机器人在线场景语义浓缩. 科学通报, 2013, 58(增刊 II): 112–120

Huang J P, Cong Y, Gao H W, et al. Online semantic condense video in robotic scene via dictional selection (in Chinese). Chin Sci Bull (Chin Ver), 2013, 58(Suppl.II): 112–120, doi: 10.1360/972013-951

了时间信息。

上述问题的实质是对视频数据进行语义浓缩,从海量数据中剔除干扰和冗余信息,只保留具有最佳语义信息的关键帧。为此,受字典选择方法^[7]和在线度量学习方法^[8]的启发,从原始视频中提取关键帧的过程可看作字典选择的过程。一般情况下,监控视频的前景帧中只有少部分的信息是非冗余的,因此首先使用背景建模从原始视频中提取出具有连续前景帧的视频段。若前景视频段过长则不利于字典选择方法处理,因此需要使用视频语义分割方法进行分割得到一系列子视频。对子视频的每一帧分别提取特征,其集合构成原始字典,最后采用字典选择方法提取出关键帧。同时,其模型也不局限于视频数据,可以用于激光、语音、文本等数据。

综上所述,本文的创新点在于:(i)我们设计了一套完整的视频语义浓缩算法,可对海量视频进行处理,同时可应用于移动机器人和固定监控平台。(ii)我们提出了一种基于 F 范数的字典选择模型,与最新的基于群稀疏性字典选择模型相比,该模型最大的优势是具有闭式解。(iii)建立新的视频数据库并进行人工标注,包含移动平台和固定平台视频。同时传统的方法依靠人工进行主观判断,这里我们设计了一种客观评价语义浓缩准确率的度量方式。

1 相关工作

语义浓缩,即关键帧提取,可使用光流^[9]或全局场景颜色差异^[10]。此类低级方法会挑选出不相关的关键帧。通过设计基于字典选择的浓缩方法,我们期望能够得到更好的效果。

语义浓缩可以采取静止图像的单一蒙太奇形式,如获得背景作为参照,并将其投影到前景区域中^[6],或者连续展示挑选出的关键姿态^[11]。一种交互式方法^[12]采用用户挑选的关键帧和关键点,生成一个表达物体轨迹的故事板。这些方法通常假设视频是具有极少前景的很短的片段,或者人工引导着提取关键帧。与此相反,导航/监控视频通常很长,提取关键帧也无需人力的介入。

Ji 等人^[13]使用一种完全自动且计算高效的体系来做视频的分析 and 浓缩。该技术使用运动物体检测和轨迹提取方法。在该方法中,首先依据运动物体检测方法将视频分段,然后从每一个运动物体中提取轨迹并挑选出关键帧,而关键帧连同轨迹共同表现

出视频段的特性。Ma 等人^[14]基于观察者注意力模型的通用语义浓缩框架,它不需要完整的对视频内容的语义理解,该框架发挥计算注意力模型的优点,消除在语义浓缩中进行复杂的启发式规则的需要。该方法使用音视频注意力模型的特征。Nam 等人^[15]提出一种新的语义浓缩方法,该方法产生一个原始视频序列的动态语义浓缩。通过保留其原始的时间特性(视觉活动)和基本语义信息,该技术简介地总结视频数据。该方法依赖于非线性采样,以事件为导向提取视频。在子视频中,局部采样率直接与视觉活动的数量成比例。Ngo 等人^[16]提出一种基于视频结构和视频集锦的语义浓缩方法,它强调内容的平衡和浓缩的感知特性。规范化的剪切算法用于全局最优地将视频分割为聚类。基于人类感知的运动聚焦模型用于计算关键帧和聚类的感知特性。聚类,连同计算出的关注值,形成一个类似于马尔可夫链的时间图,该时间图自然地描述了视频聚类的感知重要度,时间图流用于将相似的聚类分组为不同的场景,同时关注值被用于指导在场景中挑选合适的子视频来进行浓缩。Ekin 等人^[17]提出一个完全自动且计算效率高的框架来分析足球视频。它使用基于物体的特征,包含一些低级的足球视频处理算法,例如主要颜色区域检测、视频分段边界检测、关键帧分类等,对射门检测、裁判检测和处罚检测则使用一些高级算法。该系统输出3种类型摘要:(1)比赛中所有的慢动作片段;(2)比赛中所有的射门;(3)依据基于目标特征慢动作分类片段。前两种类型的浓缩对于快速过程使用基于情节的特征,后一种类型的浓缩包含高级语义特征。

2 方法概览

由第1节的分析,背景视频是无意义的,因此我们首先使用背景建模方法剔除背景帧,将前景视频段从原始视频中分离出来,该部分方法详见3.1节。过长的前景视频段难以进行下一步的操作,因此对每一个段前景视频使用视频语义分割方法再次分割,得到易于处理的子视频,见3.2节。原始字典由提取自子视频每一帧的特征组成(3.3节),最后使用字典选择方法从中挑选出关键帧(3.4节)。

方法流程如图1所示。下面进行详细介绍。

2.1 前景视频段提取

通过背景建模算法,视频中的每一帧图像均被

图1 本文方法流程

最顶行的灰色块表示前景视频段, 第二行中的浅色竖线表示使用视频语义分割技术将该段视频再次分割, 第三行中的深色竖线为使用字典选择方法挑选出来的关键帧, 其下的多张图片为从某段监控视频中提取出的关键帧

分离为前景部分和背景部分. 背景是视频中固定不变的部分, 因此其在图像中的比例越大, 则该帧图像的有效信息越小, 为剔除此类冗余信息, 我们使用背景建模方法将每一帧划分为前景帧或背景帧, 在后续的处理中直接跳过信息量较少的背景帧. 文献[18]对比了各种背景建模技术的效果. 导航/监控视频的背景建模缺少足够的训练时间, 因此选择在没有训练过程时表现最好的 VIBE 技术^[19]效果较好.

VIBE 方法使用 n 个样本点作为样本集, 每个点以其 RGB 值作为其特征. 度量当前点与样本点的距离可标识出当前点与样本点是否适合. 若当前点与超过基数个($p_{\min}$ 个)样本点适合, 则该点被判定为背景, 否则为前景, 如下式所示:

$$f_j = \begin{cases} 1, & \sum_{i=1}^m \|cd_i - sd_i\| \leq R, \\ 0, & \text{其他}, \end{cases} \quad (1)$$

$$p(x, y) = \begin{cases} 0, & \sum_{i=1}^n f_j \geq p_{\min}, \\ 1, & \text{其他}, \end{cases} \quad (2)$$

式中 cd_i 为当前点的第 i 维特征, 相应的 sd_i 为样本点的第 i 维特征, f_j 表示当前点与第 j 个样本点的适合度, R 为距离阈值, m 是特征的维数, $p_{\min}$ 为最小基数. $p(x, y)$ 为 0 则当前点 (x, y) 是背景点, 否则为前景点.

在载入视频第一帧时, 该模型随机选择 n 个当前点的邻域点作为样本点, 从而完成样本集的初始化. 若某点被分类为背景, 则有一定几率更新其自身和邻域点的样本集.

通过一帧图像中前景点的比例而将其划分为前景帧或背景帧, 如下式所示:

$$FN = \sum_{x=1}^w \sum_{y=1}^h p(x, y), \quad (3)$$

$$C = \begin{cases} \text{前景帧}, & FN/AN > \varepsilon, \\ \text{背景帧}, & \text{其他}, \end{cases} \quad (4)$$

式中 w 和 h 分别为图像宽和高, AN 为图像的总像素数, 即 $AN = w \times h$, ε 为前景占比阈值. 若一帧图像中前景像素占整幅图像像素的比例低于 ε 时, 该帧即为背景帧并将其从待处理的原始视频中剔除.

光线产生的阴影对背景建模的干扰很大, 受文献[20]启发, 本文使用 Ojala 等人^[21]提出的局部二值化算子(LBP)增强抗阴影效果. 基本的 LBP 定义如下:

$$LBP(x_c, y_c) = \sum_{p=0}^7 s(g_p - g_c) 2^p, \quad (5)$$

其中 g_c 为中心像素 (x_c, y_c) 的灰度值, g_p 为中心像素第 p 个邻域像素的灰度值. 函数 $s(x)$ 定义如下:

$$s(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (6)$$

本文所使用的背景建模特征包括像素的 RGB 值和 LBP 值, 因此(1)式中 $m=4$.

2.2 视频语义分割

视频浏览、检索、分类以及事件检测的第一步, 即是对视频进行时间分割. 现实世界中在空间和时间上紧密结合的连续视频序列即是一幕场景, 通过视频语义分割, 完整视频被分成不同的场景, 使其具有内容上的连续性.

视频语义分割可分为两类^[22]: 直接切割和渐变. 直接切割即从一个场景到下一个场景的瞬间过渡. 渐变则是由若干连续帧完成的场景过渡, 如淡入淡出、叠化、划接等.

文献[23]对视频语义分割所使用的特征和方法进行了总结. 视频语义分割使用亮度、彩色值或其直

方图、图像边缘、离散傅里叶变换、小波变化以及运动特征的一种或几种作为特征,此外,特征在时间和空间上的范围,特征的相似度量标准及分割决策方法,都是影响视频快照检测在具体环境下效果的重要因素。

本文参考文献[24]的方法,考虑监控视频亮度变化(如一天之中自然光的亮度变化),因此使用边缘作为检测的特征。 σ_n 表示第 n 帧的边缘像素的数量, X_n^{in} 和 X_{n-1}^{out} 分别表示进入第 n 帧和离开第 $n-1$ 帧的边缘点的数量,则边缘变化率定义为

$$ECR = \max(X_n^{\text{in}}/\sigma_n, X_{n-1}^{\text{out}}/\sigma_{n-1}), \quad (7)$$

其取值范围为[0,1]。为消除因物体运动而产生的度量误差,邻近的边缘像素不再被看作进入或离开的点,因此,在计算 ECR 之前采用基于豪斯多夫距离的全局运动进行补偿。

由连续前景的每一帧得到一条 ECR 值曲线,由其每个局部极大值将前景视频段切割成更短的子视频。

2.3 图像特征

在语义浓缩提取中,我们需要提取一些能够有效描述帧与帧之间区别的特征。Wu 等人^[25]使用结合时空一致性的突出测量方法从视频段中检测重要目标,该方法使用目标低级视觉特征(如目标的表观和运动信息)的空间一致性分离出物体背景,同时目标特征的时间一致性确保了随着时间改变而目标位置是一致的。在导航/监控视频中,即使在有前景出现的帧里,背景部分依然占据了图像的大部分,而前景部分只占很少的部分。描述整幅图像的特征,如图像的彩色直方图、边缘等,由前景部分对其造成的差异,都被图像中占较大比例的背景部分所淹没。

我们挑选3幅图像进行分析,它们分别具有静止(static)和不同前景(dynamic a/b)的场景。在原始图的彩色直方图中,尽管 s/a/b 三幅图像的内容差异非常大,但它们的直方图差别并不大,a/b 两图尽管包含不同的前景物体,但其直方图曲线与 s 相差无几。反观前景图的直方图,两幅动态图的直方图与背景完全可以区分开,a/b 之间的区别也比原始图要明显得多。为定量分析这种差异的程度,我们定义一个彩色直方图差异率:

$$H_{mse} = \sum_{i=1}^g \|a(i) - b(i)\|_2^2, \quad (8)$$

其中 $a(i)$ 表示图像 a 在颜色值 i 上的频度,通常彩色图像具有 $256 \times 3=768$ 级灰度,因此上式 $g=768$ 。我们分别计算 s, a, b 两两之间的差异率,结果如表 1 所示。

由以上分析可见,监控视频图像中背景画面所占比例较大,若从中提取诸如颜色直方图之类的特征,这些特征值都不明显。因此,我们需要从图像的前景部分提取特征。

前景区域的基本情况可由前景面积(即前景部分像素点的数量)、前景区域周长(即前景图中边缘点的数量)以及前景的彩色直方图表征。彩色直方图是对各彩色值的像素数量的统计,忽略像素位置特征。为此,我们使用分区域的前景彩色直方图,即将图像划分为 5 个区域,分别统计每一个区域的彩色直方图,如图 3 所示。彩色直方图实际上是对 RGB 3 个图像通道分别统计直方图,一般每个通道都有 256 个维度,实际使用中,高维度的效果并不比低维度的效果好,反而降低了运算速度,因此我们要对彩色直方图降维,即将每相邻 8 个彩色值所对应的像素数合并成 1 个维度。每一帧的前景面积不尽相同,因此需对该彩色直方图进行归一化处理,即对该区域的直方图除以该区域的前景面积。此外,前景部分的时间差异也能很好地表现前景的状况。前景的时间差异,即当前帧与上一帧中,前景部分变化了多少,具体来说就是进入点数量 FG_{in} 上一帧属于背景,而当前帧属于前景的点的数量;离开点数量 FG_{out} : 上一帧属于前景,而当前帧属于背景的点的数量。

视频中经常出现有人或物体从画面的一边移动到另一边,在类似状况下,以上所述的前景基本特征就无法很好地区分每一帧,这时,我们需要使用前景位置特征。前景位置特征就是前景区域跟画面中心的位置关系,从主观上,我们更愿意挑选监控对象(如人、汽车等)位于画面正中的图像作为关键帧,而不是物体刚刚进入或正要离开画面的那些帧,因此我们将画面中心作为原点,计算每个前景点到原点的距离并累加每个点的距离值得到位置特征。

表 1 原始图像彩色直方图的差异率^{a)}

	s	a	b
s	0	8.3(9.4)	8.6(10.1)
a	8.3(9.4)	0	8.2(9.5)
b	8.6(10.1)	8.2(9.5)	0

a) 表中数据为对原数据取 10 对数后的结果,括号内为前景图的差异率

当出现遮挡或物体分离等情况时, 还需要使用连通域特征, 它可以很好地表现前景中物体的数量、平均大小、是否有物体分离等情况. 设 c 为当前帧连通域数量, p 为上一帧连通域数量, $c(i)$ 为当前帧第 i 个连通域的像素数量, 则有以下特征: 连通域数量 c , 连通域平均面积 $\sum_{i=1}^c c(i)/c$ 以及连通域时间变化 $|c-p|$.

有时, 仅有同一个物体从画面中穿过或徘徊时, 以上特征并不能很好地区分出每一帧, 此时像素点的运动特征就变得很重要. 光流就是对每个像素运动的独立估计. 我们在当前帧的上一帧的前景图中选择一些容易跟踪的角点, 并跟踪这些角点在当前帧的位置, 从而获得其运动特征. 由于光流的计算有时并不可靠, 因此使用运动直方图更为鲁棒. 我们将角点的运动按照运动方向分为 8 个方向, 每个运动方向又按照运动速度的不同分为 3 级, 同时还要考虑到速度为 0 以及无法跟踪到(或者消失在画面中)的像素数量, 因此每个运动直方图有 26 个维度. 同前景彩色直方图类似, 我们也使用分区域的光流直方图并进行归一化.

由以上描述可知, 我们从每一帧图像里可以提取出 618 维特征.

2.4 字典选择

按照 3.3 节的方法, 我们从经过视频语义分割的子视频中提取每一帧的 d 维特征, k 帧的特征组成了基本集合 $B \in \mathbb{R}^{d \times k}$.

最普通的挑选关键帧方法是随机地或者均匀地从 B 中抽取 n 帧作为最优字典 B' , 但这显然无法完全地表现 B' , 而且会错过重要的帧, 也可能会包含意义普通的一般帧, 从而极大地影响对 B 的再现.

稀疏表示适合于使用更少的样本表现原始的高维样本. 通过稀疏重建, 最小尺寸的样本可再现原始样本. 由第 1 节对关键帧的定义可知, 关键帧的意义即是使用最少的图像再现原始视频, 因此可使用基于稀疏表示的字典选择方法来提取关键帧.

因此选择关键帧的问题转化为从基本集合 B 中, 通过字典选择的方法, 选择出具有最小尺寸的最优字典 $B' \in \mathbb{R}^{d \times n}$, 这样就从 k 帧视频里, 挑选出了 n 帧图像作为关键帧. 公式化描述如下:

$$\min_x \frac{1}{2} \|B - BX\|_F^2 + \alpha \|X\|_F^2. \quad (9)$$

这里 X 是稀疏系数, 且 $X \in \mathbb{R}^{k \times k}$. $\|X\|_F$ 为 F 范数. 使用下式可解出 X (该式的证明详见附录):

$$X = (B^T B + \lambda I)^{-1} B^T B. \quad (10)$$

正则化项在变量 X 上执行样本稀疏, 而且最优子集通常包含许多零行, 因此字典 B' 由对应于 $\|X_{i,:}\|_2 \neq 0$ 的所有样本 i 组成, 这里 $X_{i,:}$ 表示 X 的第 i 行. 系数 α 越大, 则 X 中所包含的零行越多. 我们由 X 计算出每一帧的分值, 即 $f_i = \|X_{i,:}\|_2$, 由大到小地从 f_i 中挑选出 n 个值, 其对应的帧即为关键帧. 实际应用中, 相邻的关键帧不可能离得太近, 因此我们限定所有关键帧之间的距离不得低于帧数距离阈值 W .

3 实验

为验证本文方法的有效性, 我们选取下述 4 种较先进的聚类算法进行测试对比.

(i) K-means 算法^[26]. 首先利用 K-means 对视频进行聚类, 在特征空间中与类中心最近的特征点所对应的图像, 即为关键帧.

(ii) K-medoids 算法^[27]. K-medoids 算法是一种典型的基于聚类的语义浓缩算法, k 是聚类个数或字典大小, 质心(d_i)则是每个聚类的中心点. 聚类目标是将原始集合分成 k 个聚类, 每个聚类都有最小的平方和误差:

$$\min_s \sum_{i=1}^K \sum_{x_j \in s_i} \|x_j - d_i\|^2. \quad (11)$$

(iii) Affinity Propagation 算法(AP)^[28]. AP 算法依次更新可用度和响应函数, 响应和可用度分别定义为

$$r(i, k) \leftarrow s(i, k) - \max_{k' \in \{1, \dots, k\}} \{a(i, k') + s(i, k')\}, \quad (12)$$

$$a(i, k') \leftarrow \min \{0, r(k, k')\} + \sum_{i' \in \{1, \dots, i\}} \max \{0, r(i', k')\}, \quad (13)$$

这里 $s(i, k)$ 是两点之间的相似度. 优先值的选择决定被检测到的模范点的数量, 通常设定为输入相似度的中值. AP 算法不需要预设聚类点数量, 因此可以通过调整优先值来获得合适数量的聚类.

(iv) Density-based Spatial Clustering of Application with Noise (DBSCAN) 算法^[28]. DBSCAN 算法是一种基于高密度联通区域的空间聚类算法. 该算法利用基于密度聚类的概念, 即要求聚类空间中的一定区域内所包含对象的数目不小于某一给定阈值. 其优点在于聚类速度快, 能有效处理噪声点和发现任意形状的空间聚类.

关键帧的选择并没有客观的标准,很多文献都是集合很多人的主观看法作为 Ground Truth(GT).为验证本文方法的有效性,本文设计一种客观的方法对语义浓缩的效果进行评价.本文在客观评价的同时,使用人工标记待评估视频段的关键帧作为 GT,并将其参与评估.参考(9)式,本文设计一种评估重构能力的方法.注意到(9)式的后半部份 $\alpha\|X\|_F^2$ 仅是字典选择中对稀疏矩阵的约束,而其他方法并没有对 X 的约束,因此仅使用该公式的前半部分来评估各种方法的重构能力,即

$$F_s = \min_X \|B - B'X\|_F^2, \quad (14)$$

其中 B 为原始特征矩阵, B' 为关键帧(或各个方法的中心点)的特征, X 为相应的稀疏矩阵.使用类似于附录的推导方法,得到

$$X = (B')^{-1}B. \quad (15)$$

针对机器人视觉导航和监控视频,我们使用以下两类视频进行评估.

(i) 我们使用手持摄像机模拟机器人行走,拍摄一段包含 13851 帧视频,人工标记关键帧作为 GT.由于字典一次能够处理的视频长度有限,因此从中挑选一段长度为 2299 帧的子视频进行测试.该段视频具有 4 帧关键帧(人工标记),为公平地比较各种方

法的重构能力,我们使用每一种方法得到相同数量的关键帧.

(ii) 此外,机器人视觉导航浓缩也可应用于视频监控,因此我们从商场的监控视频中挑选出 3 段子视频,分别进行测试.

评估分值和评估结果如表 2 所示,图 2 直观地表现出各种方法的重构能力.由图 2 和表 2 可见,本文方法的重构损失最小或与 GT 非常接近,同时远远小于其他方法,可见本文方法提取的关键帧能很好地重现视频信息.图 3 描绘了每段视频的分值曲线,并显示出各种方法的关键帧位置,由该图可见,本文方法选取了分值曲线的局部极大值所对应帧作为关键帧,与 GT 关键帧的位置十分相近,而其他方法的关键帧明显与 GT 相距较远,且在整个视频段上的分布非常不均衡.图 4 展示了移动平台视频段和一段固定平台的关键帧图片.综合来看,本文方法的关键帧之间相似程度很小,其他方法的关键帧之间相似度较大,表明其他方法挑选的关键帧仍然存在较大的信息冗余.

4 结论

由运动机器人和监控相机记录下的海量视频存在较大的信息冗余,降低了设备利用率,也影响对视

表 2 各种方法的重构能力比较

视频编号	视频分段	总帧数	关键帧数	各种方法的评估值					
				GT	本文	K-means	K-medoids	AP	DBSCAN
1	Nav_1	2299	4	20281	19002	46195	46195	46210	48101
2	Mall_1	2796	8	11384	13685	56183	48437	57106	48437
3	Mall_2	3930	12	10069	13759	57182	49920	54678	57130
4	Mall_3	1972	6	9421	10259	27071	27071	27395	361444

图 2 针对不同的视频数据,各种算法重构误差对比

横轴为视频段编号,纵轴为重构误差值(取其对数值)

图 3 各段视频的关键帧位置图

从上到下依次为表 2 中 4 段视频的分值曲线与各种方法的关键帧位置图. 分值越高, 表明其对应帧信息量越大

图 4 各种方法关键帧的图片墙展示

(a) 表 2 中的视频段 1; (b) 表 2 中的视频段 2. 图片下方为帧号

频内容的快速感知. 视频浓缩技术有利于视觉信息的传输、存储和浏览, 当前各种视频浓缩技术难以处理超长的导航/监控视频, 无法获得理想效果, 因此本文针对导航/监控视频的特点, 设计了一种语义浓缩算法. 为提取关键帧, 本文使用背景建模方法从原

始视频中分离出前景视频段, 使用视频语义分割方法将其切割成更短的子视频, 最后利用字典选择方法提取关键帧. 为评价本文方法的准确性, 我们通过评估其重建能力, 从客观上比较本文方法与其他几种最先进的语义浓缩方法的准确率.

参考文献

- 1 Sofman B. Online learning techniques for improving robot navigation in unfamiliar domains. Doctor Dissertation. Pittsburgh: Carnegie Mellon University, 2010
- 2 Girdhar Y, Dudek G. ONSUM: A system for generating online navigation summaries. In: Processing of the IEEE International Conference on IROS, 2010. 746–751
- 3 Cong Y, Yuan J S, Luo J B. Towards scalable summarization of consumer videos via sparse dictionary selection. In: Proceeding of the IEEE Transactions On Multimedia, 2012. 66–75
- 4 Yang C L, Shen J L, Peng J Y, et al. Image collection summarization via dictionary learning for sparse representation. Pattern Recogn, 2013, 46: 948–961

- 5 Lee Y J, Ghosh J, Grauman K. Discovering important people and objects for egocentric video summarization. In: Proceeding of the IEEE Conference on CVPR, 2012. 1346–1353
- 6 Aner A, Kender J R. Video summaries through mosaic-based shot and scene clustering. In: Proceedings of 7th European Conference on Computer Vision, 2006. 388–402
- 7 Krause A, Cevher V. Submodular dictionary selection for sparse representation. In: Proceedings of International Conference on Machine Learning, 2010. 567–574
- 8 Cong Y, Liu J, Yuan J S, et al. Self-supervised online metric learning with low rank constraint for scene categorization. *IEEE Trans Image Process*, 2013, 22: 3179–3191
- 9 Wolf W. Key frame selection by motion analysis. In: Proceedings of IEEE International Conference on Acoustics Speech, 1996. 1228–1231
- 10 Zhang H J, Wu J, Zhong D, et al. An integrated system for content-based video retrieval and browsing. *Pattern Recogn*, 1997, 30: 643–658
- 11 Caspi Y, Axelrod A, Matsushita Y, et al. Dynamic stills and clip trailers. *Visual Comput*, 2006, 22: 642–652
- 12 Goldman D B, Curless B, Salesin D, et al. Schematic storyboarding for video visualization and editing. *ACM Trans Graph*, 2006, 25: 862–871
- 13 Ji Z, Su Y, Qian R, et al. Surveillance video summarization based on moving object detection and trajectory extraction. In: Proceedings of the 2nd IEEE International Conference on Signal Processing Systems, 2010. 250–253
- 14 Ma Y F, Lu L, Zhang H J, et al. A user attention model for video summarization. In: Proceedings of the 10th ACM International Conference on Multimedia, 2002. 533–542
- 15 Nam J, Tewfik A H. Dynamic video summarization and visualization. In: Proceedings of the 7th ACM international conference on Multimedia (Part 2), 1999. 53–56
- 16 Ngo C W, Ma Y F, Zhang H J. Automatic video summarization by graph modeling. In: Proceedings of 9th IEEE International Conference on Computer Vision, 2003. 104–109
- 17 Ekin A, Tekalp A M, Mehrotra R. Automatic soccer video analysis and summarization. *IEEE Trans Image Process*, 2003, 12: 796–807
- 18 Brutzer S, Hoferlin B, Heidemann G. Evaluation of background subtraction techniques for video surveillance. In: IEEE Conference on CVPR, 2011. 1937–1944
- 19 Barnich O, Van Droogenbroeck M. ViBe: A powerful random technique to estimate the background in video sequences. In: IEEE Transactions on Acoustics Speech, 2009. 945–948
- 20 Heikkila M, Pietikainen M. A texture-based method for modeling the background and detecting moving objects. *IEEE Trans Pattern Anal Mach Intell*, 2006, 28: 657–662
- 21 Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell*, 2002, 24: 971–987
- 22 Smeaton A F, Over P, Doherty A R. Video shot boundary detection: Seven years of TRECVID activity. *Comput Vision Image Understanding*, 2010, 114: 411–418
- 23 Cotsaces C, Nikolaidis N, Pitas I. Video shot detection and condensed representation. A review. *IEEE Signal Process Mag*, 2006, 23: 28–37
- 24 Lienhart R. Reliable transition detection in videos: A survey and practitioner's guide. *Int J Image Graph*, 2001, 1: 469–486
- 25 Wu Y, Zheng N N, Yuan Z J, et al. Detection of salient objects with focused attention based on spatial and temporal coherence. *Chin Sci Bull*, 2011, 56: 1055–1062
- 26 Lee K C, Ho J, Yang M H, et al. Video-based face recognition using probabilistic appearance manifolds. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. 313–320
- 27 Hadi Y, Essannouni F, Thami R O H. Video summarization by k-medoid clustering. In: Proceedings of the 2006 ACM Symposium on Applied Computing. New York: ACM, 2006. 1400–1401
- 28 Frey B J, Dueck D. Clustering by passing messages between data points. *Science*, 2007, 315: 972–976
- 29 Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996. 226–231

Online semantic condense video in robotic scene via dictional selection

HUANG JiangPing^{1,2}, CONG Yang², GAO HongWei¹, TANG YanDong² & YU HaiBin²¹ College of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China;² State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China

Video data are important representatives for information storage. However, high similarity always exists among the video frames in many cases, such as robot videos, some kind of surveillance videos, and so on. This will lead to information redundancy, and further will cause problems in information storage, query, and transfer. In this paper, we present an algorithm for video semantic enrichment. This method may perceive the video content quickly and accurately by extracting the key frames from video sequences. First, we extract the continuous foreground segments from the original video sequences to remove the background frames. Then, we segment the foreground to be a series of sub video sequences by the shot boundary detection process. The features extracted from the sub video sequences constitute the original dictionary. Finally, we select the key frames with a dictionary selection process. By removing the useless information with a background model, our method can deal with arbitrary length videos by the shot boundary detection. The meaningful key frames can be extracted based on the image features and dictionary selection method. Compared with the artificial identification method and the baseline algorithms, the experiment results verify that our method can obtain more accurate key frames.

dictionary selection, semantic condense, robotic, online machine learning

doi: 10.1360/972013-951

附录

F 范数定义为 $\|X\|_F^2 = \sum_i \sum_j |X_{ij}|^2 = \text{tr}(X^T X)$, 而 $\text{tr}(A) = \sum_i A_{ii}$ 为矩阵 A 的迹. 因此我们可重写(9)式:

$$F_s = \arg \min_B \text{tr}(B - BX)^T (B - BX) + \alpha \text{tr}(X^T X), \quad (16)$$

这里 $B \in \mathbb{R}^{d \times k}$, 而 $X \in \mathbb{R}^{k \times k}$. 为解出该等式, 我们对其求导:

$$\frac{\partial F_s}{\partial X} = 0. \quad (17)$$

显然, 这是一个凸优化问题, 可以使用多项式优化方法解出来, 即

$$\frac{\partial \text{tr}(AB)}{\text{tr}(A)} = \frac{\partial \text{tr}(AB)}{\text{tr}(B)} = B. \quad (18)$$

因此可得

$$\frac{\partial F_s}{\partial X} = \frac{\partial \text{tr}((B - BX)^T (B - BX)) + \alpha \text{tr}(X^T X)}{\partial X} = \frac{\partial \text{tr}(B^T B - B^T BX - X^T B^T B + X^T B^T BX)}{\partial X} + \frac{\alpha \text{tr}(X^T X)}{\partial X} = 0. \quad (19)$$

则有

$$-2B^T B + 2B^T BX + \alpha 2X = 0, \quad (20)$$

即

$$(B^T B + \alpha I)X = B^T B. \quad (21)$$

此处 $I \in \mathbb{R}^{k \times k}$ 为单位阵. 一般来说, $\alpha > 0$, 因此 $(B^T B + \alpha I)$ 是满秩矩阵, 其逆矩阵存在, 于是得 X 的闭式解

$$X = (B^T B + \alpha I)^{-1} B^T B. \quad (22)$$