基于内容代表性评价的关键帧抽取

顾益军[1] 解 易[1] 夏 天[2,3]
(中国人民公安大学信息安全保卫学院 北京 100872)[1]
(中国人民大学数据工程与知识工程教育部重点实验室 北京 100872)[2]
(中国人民大学信息资源管理学院 北京 100872)[3]

摘 要 视频关键帧提取技术是对视频进行摘要来提高视频内容访问效率的一种操作。传统的方法主要采用聚类的方法,未给出可信的关键帧代表性描述。尝试基于图计算算法实现关键帧抽取,该算法可以将一段视频中候选帧及其之间的关系表示成一个相关图,通过各帧间基于相关性对相邻帧的分值分配进行迭代计算,实现候选帧内容代表性评价;并提出了一种高效的帧间相关性计算方法。该方法通过两帧图像的最大稳定颜色区域(maximally stable colour region,MSCR)的匹配情况判定它们的相关性。在测试视频上将该算法与传统算法进行了对比测试,测试的结果验证了该算法的有效性。

关键词 关键帧提取,相关性计算,视频
中图法分类号 TP391,G358 **文献标识码** A DOI 10.11896/j.issn.1002-137X.2014.08.060

Keyframe Extraction Based on Representative Evaluation of Contents

GU Yi-jun[1] XIE Yi[1] XIA Tian[2,3]
(Schools of CyberSecurity,Chinese People's Public Security University,Beijing 100872,China)[1]
(MOE Key Laboratory of Data Engineering and Knowledge Engineering,Renmin University of China,Beijing 100872,China)[2]
(School of Information Resource Management,Renmin University of China,Beijing 100872,China)[3]

Abstract The keyframe extraction is a visual summary method. It enhances the accessibility to the visual content. Traditional methods extract keyframes through clustering. These methods don't provide reliable descriptions of keyframe representative. This paper proposed a novel keyframe extraction method through a graph model representing the candidate keyframes and the correlations between them. The representative of candidate keyframe was calculated through propagating grade between correlated candidate keyframes iteratively. To support the calculation of the representative, the paper introduced a correlation calculation method according to how well the maximally stable colour regions of two frames match to each other. The experiments were conducted on several test videos and the results validated our keyframe extraction method.

Keywords Keyframe extraction,Correlation calculation,Video

1 引言

随着数字多媒体技术的发展,人们通过各种渠道接触到的视频内容总量迅速增加。为了能够让人们可以更加便捷地访问所需的多媒体内容,数字媒体摘要技术成为研究的热点。关键帧抽取是视频内容摘要的一个重要分支,其目标是通过处理而得到视频的一个紧凑的表示,且这个表示应能表达原视频中表现的重要活动。

针对关键帧抽取问题人们已经开展了大量研究,最常见的算法是通过聚类获得镜头片段,并从每个镜头片段中获得关键帧。Zhuang 等人在文献[1]中通过一种带粒度回馈的无监督聚类方法将视频中所有帧聚类成为镜头片段,并用聚类中心作为每个镜头片段的关键帧。Zhao 等人在文献[2]中通过最近特征线(nearest feature line)方法将视频聚类成镜头,然后在镜头中寻找关键帧。许多算法在聚类提取关键帧基础上进行了改进,抛弃简单的采用聚类中心作为关键帧的策略,而选择设置更加合理的关键帧提取条件。Pan 等人在文献[3]中同样通过聚类的方法划分视频的镜头和子镜头片段,然后在每个聚类中寻找最大熵的帧作为关键帧。Liu 等人在文献[4]中在每个镜头片段中通过运动模式来寻找关键帧。Li 等人在文献[5]中通过动作关注方法在每个镜头中寻找关键帧。无论如何,这些方法都没有离开聚类的框架。本文参照网页链接分析中的 PageRank[6] 算法,提出一种基于图计算的关键帧提取方法。本文算法定义帧代表性为各帧基于内容相

到稿日期:2013-06-03 返修日期:2013-07-26 本文受公安部重点研究计划项目(2011ZDYJGADX016),北京高等学校青年英才计划项目(YETP1366)资助。

顾益军(1968—),男,博士,副教授,主要研究方向为网络情报获取分析技术,E-mail:yj_gu@163.com;解 易(1984—),男,博士,讲师,主要研究方向为网络情报技术;夏 天(1978—),男,博士,副教授,主要研究方向为信息检索、Web 数据挖掘。

似性和位置差异进行影响力转移的结果。这种方法与聚类划分镜头片段的方法相比更注重选取较少的帧，实现对于整段视频的“鸟瞰”，从而提高视频检索效率。

为了配合关键帧提取算法，本文还提出了一种新的帧间相关性的度量方式。因为视频中场景和目标的位置在两帧之间一般不会完全相同，所以相关性不可以用像素级别的匹配方法。本文提出了一种通过图像的最大稳定颜色区域(MSCR)[7]匹配情况估计相关性的方法。这种匹配方法淡化了场景和目标的轻微位置变动，成为一种更加可靠的帧间相关性度量方法。

2 帧的相关性计算

2.1 相似度计算思路

为了给本文提出的关键帧提取算法提供帧间相关度信息，我们提出一种采用最大稳定颜色区域匹配的方法。最大稳定颜色区域(MSCR)[7]是最大稳定极值区域(MSER)[8]的颜色拓展。最大稳定极值区域最开始是 Matas 等人在文献[8]中提出的一种从不同视角拍摄相同场景的两幅图像中寻找对应元素的技术。极值区是一些像素区域，这些像素区域内部的所有像素灰度都大于或者小于区域边界灰度。极值区可以通过添加符合要求的像素扩充成更大的极值区。扩充时，如果一个极值区只需要添加非常少的像素就成为另一个极值区，那么这个极值区为最大稳定极值区。研究发现，最大稳定极值区具有仿射不变性，所以灰度图像中的最大稳定极值区可以作为图像匹配的有效依据。最大稳定颜色区域是最大稳定极值区的拓展，它在考虑灰度关系的同时，还考虑了颜色的相似度。我们采用匹配最大稳定颜色区域来确定两帧图像的内容相似度。

2.2 相似度计算过程

比较相关度的第一步是在两张图像 I_x 和 I_y 中提取最大稳定颜色区域。然后通过宽基线算法(wide baseline algorithm)[8]匹配两幅图中的这些区域。宽基线匹配算法是一种立体视觉和运动匹配中常用的图像基原(image primitive)匹配算法。该算法通过各种特征点提取算法在两幅图像中寻找能够进行匹配的特征点。因为并不是所有特征点之间的匹配都是可靠的，所以采用 RANSAC 算法[9]估计这些特征点匹配中的外点，并用内点估计两幅图像之间的仿射关系。

在我们的算法中，采用最大稳定区域中心点作为特征点进行匹配。如果得到的最大稳定颜色区域的匹配少于 3 个，则定义两幅图像的相关度为 0。如果匹配的最大稳定颜色区域对超过 3 个，那么就用这些最大稳定颜色区域的重心位置通过 RANSAC 算法[9]估计一个仿射矩阵 A_{xy}。然后通过

$$L(I_x,I_y)=\exp\{-\sum_{k=1}^{3}\ln^2|\lambda_k(A_{xy})|\} \tag{1}$$

计算两幅图像的内容相似度。其中 $\lambda_k(A)$是矩阵 A 的第 k 个特征值。当两幅图像之间的仿射变换几乎为单位阵时，其内容相似度非常高。若两幅图像是从完全不同的镜头拍摄的，那么两幅图内容相似度非常低。RANSAC 算法保证了即使部分最大稳定颜色区域是移动的前景，也不会影响到两帧图像场景的仿射矩阵的估计。

在不同的镜头中可能会重复出现相似图像。比如新闻联播中，一段新闻报道的开头和结尾都会出现新闻主播播报新闻的画面，显然它们应该属于不同的镜头，但因为新闻主播的姿势和拍摄的角度保持不变，所以仅仅考虑两帧图像是否相似就确定它们的相关度还是不够的，在计算相关度的时候还应考虑到两帧图像在视频中的相对位置。两帧图像如果帧号相差很远，那么很有可能来自两个镜头，所以相关度应当相应降低，以客观表达两帧图像间的真实相关性。反之，相关度应该近似等于内容相似度。根据上述考虑，相关度通过

$$s(I_x,I_y)=L(I_x,I_y)\cdot\exp\{-|x-y|/\Delta\} \tag{2}$$

进行计算。其中 x 和 y 是图像 I_x 和 I_y 对应的帧号。Δ 是后面将要介绍的图像采样间隔。通过式(2)计算两帧非常相似的图像的相关度时，如果它们之间的距离较远，相关度仍然会相应降低。

3 关键帧抽取

3.1 候选帧的关系图表示

在对实际的视频提取关键帧的时候，视频中包含很多帧图像，出于效率考虑，关键帧仅在视频的候选帧中进行选择。候选帧通常是以用户设定的 Δ 为采样间隔，在图像中提取的采样帧集合。如果将两个候选帧之间的相关度看作是候选帧的关系，我们可以将一段视频看作候选帧的关系图 $G(T,E)$，其中以候选帧集构成节点集 T，以各帧间相关性构成边集 E。图 1 为一个候选帧的关系图表示示例。

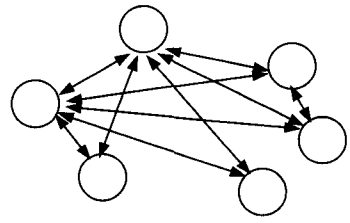

图 1 候选帧的关系图表示示例

3.2 基于代表性评价的关键帧抽取

关键帧抽取首先需要解决的是每个候选帧的内容代表性评价，借鉴 PageRank[6]的思想，可以构造一个基于关联关系强弱所构成的随机游走模型，实现内容代表性的赋值传递，以完成候选帧节点集内容代表性评价排序。

根据候选帧集构成节点集中元素个数 k，初始化候选帧的代表性向量 V，向量中的每个分量 v_i 为节点 t_i 的初始化权重 $\frac{1}{k}$，则 $V=V^{(0)}=(\frac{1}{k}\ \frac{1}{k}\ \cdots\ \frac{1}{k})$。

基于式(2)得出的节点 i 和节点 j 的相关度 $s(I_i,I_j)$，计算 $s(I_i,I_j)$针对 i 相邻节点相关度之和的归一化取值，可以得出节点 i 对节点 j 的权重分配系数：

$$a_{ij}=s(I_i,I_j)/\sum_{k:v_i\to v_k}s(I_i,I_k) \tag{3}$$

图 2 为一个节点权重分配示例。

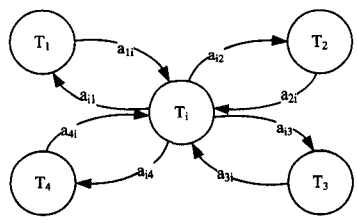

图 2 节点权重分配示例

基于节点间权重分配系数，可以得出候选帧关系图的邻接矩阵为

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \tag{4}$$

通过迭代公式

$$V^{i+1}=\beta V^{i}A+(1-\beta)\frac{1}{|v|},i\geqslant 0 \tag{5}$$

迭代至 $V^{(i)}$ 收敛后即为候选帧的代表性。其中 β 称为缓冲因子，$1-\beta$ 为各节点的随机游走概率，用来保证存在非强连通关系图时，同样可以收敛。

在得到候选帧的代表性后，我们将代表性最高的候选帧选为关键帧。将所有的候选帧按照其代表性降序排序，然后选择帧间相关度不大于 θ 的前 n 帧作为关键帧。

4 实验

4.1 实验数据介绍

实验中的视频采自与航天相关的新闻的片段。新闻因为需要通过画面说话，所以会采用蒙太奇效果衔接多个画面。我们采用的新闻片段中存在丰富的画面衔接，有些画面之间存在内容的相关性，有些画面之间相关性相对较弱。在这些视频上测试我们的关键帧算法，以验证算法能否从相关性的角度提取出较强代表性的关键帧，而不受画面切换的影响。

4.2 定性结果

本文提出的基于内容代表性评价的关键帧提取算法能够提取出整段视频中极具代表性的图像。本文提出的算法重点在能够自动地提取更少的最具代表性的关键帧，所以为了对比基于聚类的关键帧提取算法，我们将 K-mean 聚类关键帧提取算法的结果与本文所提算法的结果进行了对比，通过对比说明我们的算法在网络视频监控应用中的优势。

视频 1 是一段宇航飞船的 3D 模拟视频。它包含近 7000 帧、44 个镜头。通过 K-mean 聚类关键帧提取算法提取的关键帧如图 3 所示。通过代表性关键帧算法提取的关键帧如图 4 所示。

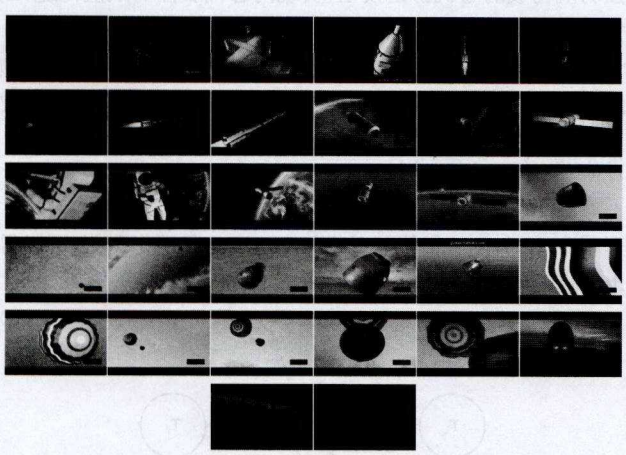

图 3 K-mean 在宇航飞船 3D 模拟视频上提取的关键帧

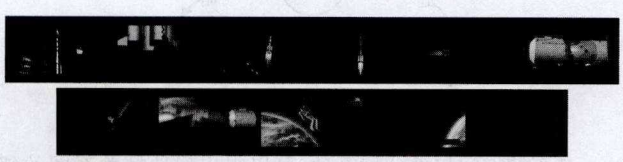

图 4 本文算法在宇航飞船 3D 模拟视频上提取的关键帧

从实验结果可以发现，我们的算法提取出了更少的关键帧。这是因为，我们的算法并不基于聚类的分镜头方法，而是通过 PageRank 算法找到与相邻帧相关度较大的帧。按照本文提出的最大稳定颜色区域的帧间相关度计算方法，我们的算法提取的所有关键帧都处在画面比较稳定的时候，我们认为这样的画面才可以代表视频的主要内容。所以我们的算法忽略了所有画面变化频繁的片断，尤其是这段视频的最后，神舟号落地阶段，画面变动较大，只在最后落地后提取了一帧关键帧。从代表性来看，我们的算法可以用更精简的关键帧完整地反映视频的内容。

视频 2 是一段宇航员出舱讲话的视频。它包含近 7500 帧、12 个镜头。通过 K-mean 聚类关键帧提取算法提取的关键帧如图 5 所示。通过代表性关键帧算法提取的关键帧如图 6 所示。

图 5 K-mean 在宇航员出舱讲话视频上提取的关键帧

图 6 本文算法在宇航员出舱讲话视频上提取的关键帧

这段视频中存在大量的快速变化的渐变镜头，稳定的镜头不多。所以如果采用聚类分镜头方法提取关键帧就会产生大量的内容重复的关键帧。我们看到，K-mean 算法在视频中提取了内容相似的帧，本文算法只选择了 3 帧在画面稳定的时候具有代表性的关键帧。这是本文算法在这段视频中提取关键帧的优势展现。

视频 3 是一段神舟七号的新闻报道，它包含近 1500 帧、10 个镜头。通过 K-mean 聚类关键帧提取算法提取的关键帧如图 7 所示。通过代表性关键帧算法提取的关键帧如图 8 所示。

图 7 K-mean 在神舟七号的新闻报道视频上提取的关键帧

图 8 本文算法在神舟七号的新闻报道视频上提取的关键帧

新闻报道中经常出现短镜头。对于酒泉卫星发生中心的展现需要多组镜头不断进行切换，所以 K-mean 聚类关键帧提取算法提取出了很多对应多个镜头的关键帧，我们的算法提取出了较少的能够对这个新闻报道起到摘要作用的关键帧。

视频 4 是另一段神舟七号的新闻报道，它包含近 1200

（下转第 315 页）

[8] Yang J, Yan S, Fu Y, et al. Non-negative graph embedding [C]// Proceedings of IEEE International conference on Computer Vision and Pattern Recognition. 2008:1-8

[9] Yan S, Xu D, Zhang B, et al. Graph embedding and extensions: a general framework for dimensionality reduction [J]. IEEE Trans. on Pattern analysis and machine intelligence, 2007, 29(1):40-51

[10] Guan N, Tao D, Luo Z, et al. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent [J]. IEEE Trans. on Image Precessing, 2011, 20(7):2030-2048

[11] Chen H T, Chang H W, Liu T L. Local discriminant embedding and its variants [J]. Proceedings of IEEE International Conference on Computer Vision and Patern Recognition, 2005(2):846-853

[12] Ding C, Li T, Jordan M I. Convex and semi-nonnegative matrix factorizations [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010, 32(1):45-55

[13] 方蔚涛,马鹏,成正斌,等. 二维投影非负矩阵分解算法及其在人脸识别中的应用[J]. 自动化学报,2012,38(9):1503-1512

[14] 刘雪松,王斌,张立明. 基于非负矩阵分解的高光谱遥感图像混合像元分解[J]. 红外与毫米波学报,2011,30(1):1-5

[15] 于红芸,姜涛,关键. SAR 图像目标检测的互信息非负矩阵分解算法[J]. 中国图象图形学报,2011,16(1):129-134

[16] Yang Z, Oja E. Linear and nonlinear projective nonnegative matrix factorization [J]. IEEE Trans. on Neural Networks, 2010, 21(5):734-749

[17] Lee D D, Seung H S. Algorithms for non-negative matrxi factorization [J]. Advances in Neural Information Processing Systems, MIT Press, 2001, 13:556-562

[18] Ding C, Li T, Peng W, et al. Orthogonal nonnegative matrix trifactorizations for clustering [C] // Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006:126-135

[19] Georghiades A S, Belhumeur P N, Kriegman D J, et al. From few to many: illumination cone models for face recognition under variable lighting and pose [J]. IEEE Trans. Pattern Anal. Mach. Intelligence, 2001, 23(6):643-660

[20] Sim T, Baker S, Bsat M. The CMU pose, illumination, and exprssion database[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2003, 25(12):1615-1618

(上接第 288 页)

帧、16 个镜头。通过 K-mean 聚类关键帧提取算法提取的关键帧如图 9 所示。通过代表性关键帧算法提取的关键帧如图 10 所示。

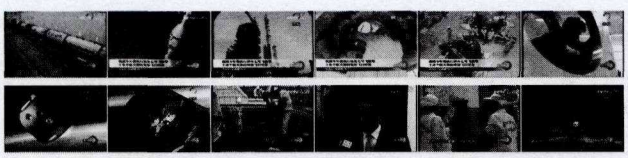

图 9 K-mean 在神舟七号的新闻报道视频上提取的关键帧

图 10 本文算法在神舟七号的新闻报道视频上提取的关键帧

这段视频同样是富含短镜头的新闻报道。在这段视频中,我们的算法可以提取出更精简的极具代表性的关键帧。

从实验中我们可以看到,我们的算法不拘泥于按照镜头提取关键帧的方法,而是直接针对画面的代表性提取关键帧。这使我们的算法提取的关键帧不会随着视频的镜头数量增加而增加,使提取的关键帧具有真正代表视频内容的功能。在网络视频监控应用中,海量的视频信息需要通过最有效的方式进行摘要。这种需求使本文提出的算法可以在网监应用中更好地发挥作用。

结束语 我们在这篇文章中介绍了一种通过计算帧的代表性并基于代表性评价而实现的关键帧提取算法。在计算代表性的过程中我们提出了一种计算帧间相关度的算法,并利用了一种改进的 PageRank 算法完成代表性的计算。通过与传统的基于聚类的关键帧提取算法在 4 段视频上获得的关键帧进行对比,验证了我们的算法能够通过更少的极具代表性的关键帧对视频进行摘要。实验结果验证了算法的有效性。

参 考 文 献

[1] Zhuang Y, Rui Y, Huang T S. Video key frame extraction by unsupervised clustering and feedback adjustment[J]. Journal of Computer Science and Technology, 1999, 14(3):283-287

[2] Zhao L, Qi W, Li S Z, et al. Key-frame extraction and shot retrieval using nearest feature line (NFL)[C]// Proceedings of the 2000 ACM workshops on Multimedia. New York: ACM, 2000: 217-220

[3] Pan R, Tian Y, Wang Z. Key-Frame Extraction Algorithm Based on Entropy[C]// 2010 International Conference on E-Product E-Service and E-Entertainment (ICEEE). IEEE, 2010:1-4

[4] Liu T, Zhang H J, Qi F. A novel video key-frame-extraction algorithm based on perceived motion energy model [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(10):1006-1013

[5] Li C, Wu Y T, Yu S S, et al. Motion-focusing key frame extraction and video summarization for lane surveillance system[C]// 2009 16th IEEE International Conference on Image Processing (ICIP). Cairo: IEEE, 2009:4329-4332

[6] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks and ISDN Systems, 1998, 30(1):107-117

[7] Forssen P E. Maximally stable colour regions for recognition and matching[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2007(CVPR'07). IEEE, 2007:1-8

[8] Matas J, Chum O, Urban M, et al. Robust wide-baseline stereo from maximally stable extremal regions[J]. Image and Vision Computing, 2004, 22(10):761-767

[9] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6): 381-395