基于时空轨迹行为特征的视频拷贝检测算法

吴潇[1,2]，李锦涛[1]，吴宇锋[3]，郭俊波[1]，任化敏[1,4]

（1.中国科学院计算计算研究所，智能信息处理重点实验室，北京 100190）；

(2.中国科学院研究生院，北京 100085)；

(3.大连理工大学，自动化技术研究所，大连 116023)；

(4. 北京中医药大学，信息中心，北京 100029）

摘 要： 互联网环境中大规模的视频拷贝检测面临两个难题，拷贝变化多样性问题和数据稀疏问题。有效的拷贝检测算法必须使用鲁棒、快速、精简的视觉特征来应对各种拷贝变化。本文提出利用视频连续帧的特征点轨迹的行为，来构造视频内容的不变模式特征；并使用轨迹视觉关键词典进行快速的拷贝定位，解决数据稀疏问题。在标准数据集上的对比实验证明，本文提出的算法在各种常见的拷贝变化下可以得到更好的检测精度，其特征提取的时空复杂度更低，适合大规模数据的实时拷贝检测。

关键词： 视频拷贝检测；时空视觉轨迹；局部特征点检测；时序一致性匹配；

Video Copy Detection Based on Spatio-Temporal Trajectory Behavior Feature

Xiao Wu[1, 2], Jintao Li[1], Yufeng Wu[3], Junbo Guo[1], Huamin Ren[1,4]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190) 、

(2. Graduate School of the Chinese Academy of Sciences, Beijing 100085) 、

(3. Institute of Automation, Dalian University of Technology, Dalian 116023) 、

(4. Information Center, Beijing University of Chinese Medicine, Beijng 100029)

Abstract: Large scale video copy detection task requires compact feature insensitive to various copy changes. Based on local feature trajectory behavior we discover invariant visual patterns for generating robust feature. Bag of Trajectory (BoT) technical is adopted for fast pattern matching. Our algorithm with lower cost is more robust compared to the state-of-art schemes.。

key words: video copy detection; spatio-temporal visual trajectory; local feature point detection; matching based on temporal info

1. 前言

随着网络带宽的增长和视频编辑软件的普及，越来越多的用户参与互联网的多媒体内容制作，比如将自己制作的视频上传到分享网站上。网络视频的兴起带来了诸多问题，如有版权的视频内容被非法截取修改等。研究者提出使用基于内容的视频拷贝检测技术来解决这类问题。视频拷贝检测的任务是，设计算法由计算机自动判断两个不同的数字视频是否存在共同的片段（拷贝）。在互联网上，视频拷贝常常被施加了各种拷贝变化，常见的图像变化如改变分辨率，亮度和插入字幕标志等；视频编码变化如帧率，码率改变等。拷贝变化使得视频拷贝段和原始的片段视觉上有很大不同[1]，无法使用相似性检索的算法进行检测。所以如何发现视频内容的不变模式特征对拷贝检测的研究具有重要意义。另外，包含拷贝片段的视频不多，拷贝片段的时长对于整个视

基金资助：本研究由国家重点基础研究发展计划项目(973 项目， 2007CB311100)，国家高技术研究发展计划项目(863 项目，2007AA01Z416)，国家自然科学基金项目(60773056)和北京市科技新星计划项目(2007B071)共同资助。

作者简介：吴潇（1982-），男，湖南省株洲市人，中国科学院计算技术研究所在读博士生，本科毕业于湖南湘潭大学信息工程学院计算机系。Email: wuxiao@ict.ac.cn

频而言也是较短。这类现象可以归结为拷贝稀疏问题，需要使用高效的拷贝定位技术找到这些片段。本文中我们挖掘视频内容的不变视觉模式，研究适用的拷贝检测算法以应对以上提到的问题。

我们的研究工作旨在利用视频的时空信息，建模视频内容的不变视觉模式，作为拷贝检测的特征。在第三节，我们介绍如何使用 Harris 局部不变特征点检测算子结合快速 KLT 点跟踪算法，提取视频连续帧的鲁棒特征点轨迹。在提取轨迹的同时，视频被分割为大量亚镜头（sub-shot），我们在亚镜头级别上进行拷贝定位。在第四节中，通过使用轨迹关键词技术，每个亚镜头被表示为特征词频向量，拷贝检测问题被转化为模式匹配问题进行求解。第五节的实验证明我们的算法比国际先进技术更加鲁棒和高效，并产生精简的描述特征，适用于大规模网络视频拷贝检测。

2. 相关工作

视觉特征提取是视频拷贝检测最重要的因素。理想的视觉特征必须具备以下三点要求：1），鲁棒性：对各种常见的拷贝变化不敏感；2），精简性：占用少量的磁盘空间；3），高效性：提取过程的计算复杂度较低。

一类前人的相关研究[1,2,3,6,7,8]专注于研究如何提取各种全局、局部的视频帧图像描述特征。其共同点是在帧级别进行视频内容描述。比如早期的一些研究[2][3]提出使用精简的图像数字签名作为特征，但这些签名一般基于简单的图像统计特征，如边缘统计直方图，对一些图像全局变化和大部分局部变化都非常敏感。基于图像分块序列的 OM 方法（Ordinal Measure）[2,3,8]通过发现图像块间相对关系来构造不变视觉特征。但是局部变化一般会打乱图像块间的相对关系，使得这类方法失效。近年来，在计算机视觉领域研究成熟的图像局部特征点检测算子[4]和描述算子[5]对拷贝变化鲁棒，被用于拷贝检测研究中[6][7]。但是大都过于追求完美匹配，计算量和存储量特别大，没有考虑到网络视频数据的稀疏问题和实时性检测要求。

另一大类研究利用视频序列的时序信息构造特征做视频片段的配准[9]。这类方法利用的时序关系对视觉变化有一定的鲁棒性，却对时域的变化非常敏感，如前后片段调换，帧率降低和掉帧等。其实验假定进行匹配的两段视频等长，或者视频间必然有公共片段。这类前提不符合视频拷贝检测的稀疏现象。

针对拷贝变化多样性和数据稀疏问题，关键研究点是提取鲁棒精简的不变特征。不同于先前的研究，我们使用包含时空信息的轨迹行为作为特征，以亚镜头为单位进行拷贝定位。

3. 构造不变视觉模式

我们基于视频的帧图像局部特征点的轨迹来构造对拷贝变化鲁棒的不变视觉模式。视频物体的轨迹在视频监控，物体跟踪等其他研究中有很多应用。近年来局部特征点[5]轨迹更是在视频内容分析的研究[6][11][13]中被深入研究。本文算法首先提取稳定的视频轨迹，然后提取其行为特征作为不变视觉模式用来匹配亚镜头。

3.1 快速提取稳定的视频轨迹

特征点检测和跟踪是提取连续帧轨迹的两大技术要点。Harris 特征点检测算子作为最快速的局部特征点检测算法[4]，已经被用于拷贝检测的相关研究中[6]，被证明具有一定的鲁棒性。然而每帧都检测 Harris 特征点将耗费较大的计算量，并较其他检测算子产生更多的特征点[4]使得后续工作量很大。[6]对大量的特征点每个点计算 20 维的区域边缘直方图特征，匹配起来复杂而冗余。KLT[12]也在最近的研究[13]中被用于拷贝检测。[13]使用的是 KLT 本身的特征点检测算法，特征点数目必须固定，不适合变化的视频内容。

图 1 使用 Harris 和 KLT 算法产生的视频轨迹，黑白帧表示亚镜头分割处即轨迹断开处

Fig.1 Generating local feature trajectories using Harris detector and KLT tracker

本研究中，我们提出使用 Harris 算子结合 KLT 特征点跟踪算法快速提取视频轨迹。即利用 Harris 检测第一个视频帧的所有特征点，然后使用 KLT 跟踪算法在后续帧跟踪这些特征点。当点数目突然下降（如镜头场景切换的原因）或者点数少于一定阈值（如跟踪过程中会丢失一些点）的时候，重新检测下一帧的 Harris 特征点并跟踪，如图 1 所示形成新的轨迹簇。我们把轨迹簇断开处作为边界，称边界之间的视频子段为亚镜头，以亚镜头为单位提取轨迹行为特征和做匹配定位。

3.2 建模轨迹行为模式

视频轨迹包含了帧图像局部特征点在连续帧的运动行为信息，而这些特征点大部分能在各种拷贝变化下存在，从而保证拷贝检测的精度。为了取得精简的轨迹描述，本文提出一种简单的编码算法建模轨迹行为，如图 2 所示，轨迹上的特征点相对位置的变化被编码为 5 种代码，一条轨迹可以被表示一个编码序列。

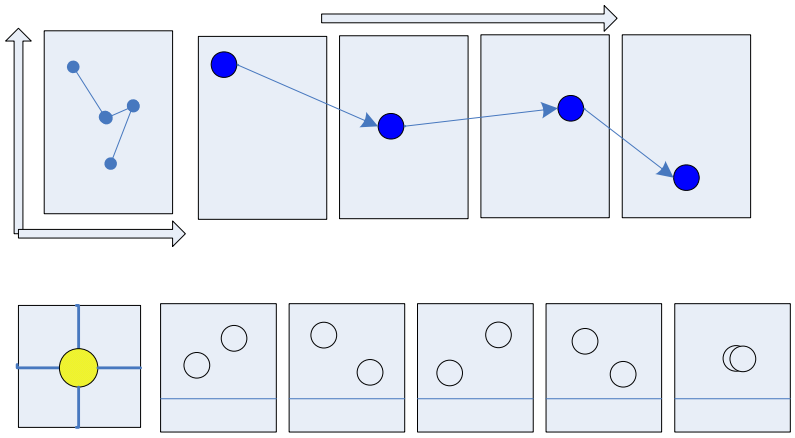


图 2 （a）连续帧的对应点的位置变化 （b）特征点在各自的帧中的位置 （c）四象限加中心区域的编码模板，用于建模轨迹行为特征，中心区域的大小定义了静态特征点的运动范围，设定为 7×7 大小 （d）特征点的位置关系和对应的编码数字

Fig.2 (a) Relative position of a point on continual frames (b) feature point “walk” on a video trajectory (c) quadrant template for point behavior encoding, the center area defines the scale of point repeatability (d) position relations and the corresponding codes

单个的代码存储了轨迹的空间信息，代码的顺序建模了轨迹的时间信息。如 3.1 小节所述，由于 Harris 检测算子和 KLT 跟踪算法的鲁棒性，常见的全局拷贝变化一般只会有轻微影响轨迹的数目和长度。但局部变化，如遮挡等，和时域的拷贝变化，如丢失帧，帧率降低等，会改变轨迹的数目和亚镜头分割的位置（即改变了轨迹的长度），进而改变其行为模式。针对这些问题，我们提出一种简单的特征提取方法，通过把轨迹行为编码数列进行归一化统计，生成 5 维直方图，平滑掉各种变化带来的影响。如式（1）所示，

$$CH(i) = \#i / Length(Code\_Sequence) \tag{1}$$

举例来说，序列{4，1，3}被转换的直方图为，{0:0, 1:0.33, 2:0, 3:0.33, 4:0.33}。轨迹越长，直方图的区分性越强。我们使用这种简单的转换，保留轨迹行为的统计信息，消除了大部分各种变化可能带来的噪声影响，并方便轨迹之间比较和轨迹聚类。

4. 匹配不变视觉模式

经过提取视频不变视觉模式（即轨迹），一个视频被分割为若干的亚镜头。每个亚镜头都包含一组轨迹特征直方图。类似于先前在图像检索中使用的视觉关键词技术[14]（Bag of Visual Words，BoW），我们把亚镜头看做包含一组轨迹关键词的文档（Bag of Trajectory，BoT）。本节中，基于这类视觉关键词的思想提出不变视觉模式匹配算法，用于拷贝定位。

4.1 离线计算：生成轨迹关键词典

类似于视觉关键词技术[14]，在查询之前，我们将源视频数据库中的所有视频提取轨迹，建模为大量的 5 维直方图特征，然后利用经典Kmeans算法进行聚类，产生一个轨迹行为特征词典，每个关键词都是一个有代表性的 5 维特征。Kmeans把大量轨迹行为聚类到少量的类别上，去除了数据冗余并削弱了噪声的影响。假设生成K个聚类中心作为关键词，每个亚镜头可以表示为K维词频向量，$V_s=\{t_1, t_2, \ldots, t_i, \ldots, t_k\}$，$t_i$表示轨迹特征关键词i在亚镜头s中出现的次数。但是词典的规模K必须人为指定，K的设置对聚类效果有很大影响，对于本研究使用的数据集我们在测试不同K值下的检测精度（见图 4（a）），得到一个适合此数据集的最优K值。

在实验中也观察到，静态镜头中大量的特征点保持位置不变，其特征数列为全 0 序列，一些短轨迹叶缺乏区别性，这类在文本检索和图像检索中共有的现象称为停用词现象（stop words）。参考文本检索的经典思想，我们利用 tf-idf 公式对亚镜头词频向量做加权操作，如式（2）所示，

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i} \tag{2}$$

其中n_{id}是关键词i在亚镜头d中出现次数，n_d是亚镜头d包含的关键词个数。n_i表示关键词i在所有亚镜头中出现的次数，而N表示词典的大小即K。使用tf-idf加权后，停用词由于其n_i数目大，自动降低了其权值和影响力，达到了去停用词的目的。

4.2 在线计算：基于亚镜头的拷贝匹配和定位

在线查询过程中，查询视频首先被提取轨迹并分割为若干亚镜头（如 3.1 小节所述）。然后被计算 5 维特征并由关键词词典归类为关键词，每个查询亚镜头被表示为词频向量。

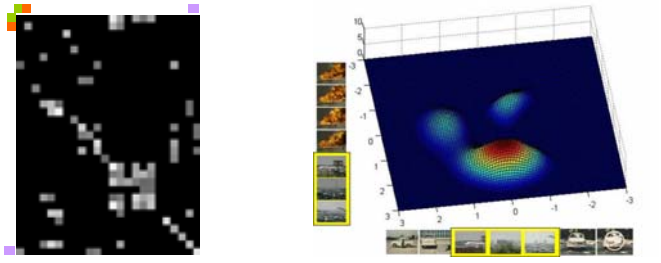
4.2.1 使用余弦聚类进行亚镜头匹配

查询亚镜头和库视频亚镜头的相似度的计算可以使用多种方法。我们在本研究中使用计算复杂度很低的余弦距离来计算亚镜头相似度，如式子（3）中所示。全部 Sim 组成相似度矩阵，如图 3（a）所示。

$$Sim(SS_a, SS_b) = Cosine\_Sim(V_a, V_b) \tag{3}$$

4.2.2 基于亚镜头相似度矩阵的视频拷贝定位

拷贝检测的目的是要定位查询视频与源视频共有片段的起始位置。基于亚镜头相似度矩阵，除了正确的匹配块结构，还存在很多零散的错匹配。同时，由于拷贝的稀疏性特点，查询视频常常不包含拷贝片段。为了去除这些噪声，我们在矩阵上使用分水岭算法使用一个阈值作为水平面 “淹没”低相似度的错配噪声，从而发现正确匹配的位置。如图 3，在实验中正确匹配段常常表现为“山脊”和“高原”的峰值结构，我们计算这些块结构的相似度和，选择最大的部分定位拷贝。



（a） （b）

图 3 （a）亚镜头相似度矩阵的例子 （b）使用分水岭算法淹没噪声，发现匹配的“山脊”“高原”结构

Fig.3 (a) Example similarity matrix of query and referent sub-shots (b) The copy clip (airplane) is located by discovering the “ridge” using watershed algorithm. Solo “peaks” of smaller total similarities are discarded.

5. 数据集和对比实验

本文实验采用的源视频数据库是 CIVR 2007 国际拷贝检测公开评测和 TRECVID 2008 年拷贝检测比赛[16]共同使用的数据集，由法国 MUSCLE 研究部门提供[15]。查询视频的制作遵循 TRECVID 2008 年拷贝检测比赛使用的方法[16]。我们从 20 个源视频中截取小段（3 秒到 1 分钟）嵌入到 20 个 TRECVID 2007 的高层概念检测的短视频中作为拷贝查询。另外选取 10 个高层概念检测的短视频作为非拷贝查询。30 个查询视频分别施加[16]中定义的各种时域，空域拷贝变化，改变视频的画面和帧序等。

表 1 使用主流台式机进行实验数据集处理的相关信息

Tab.1 Information of video set for experiments

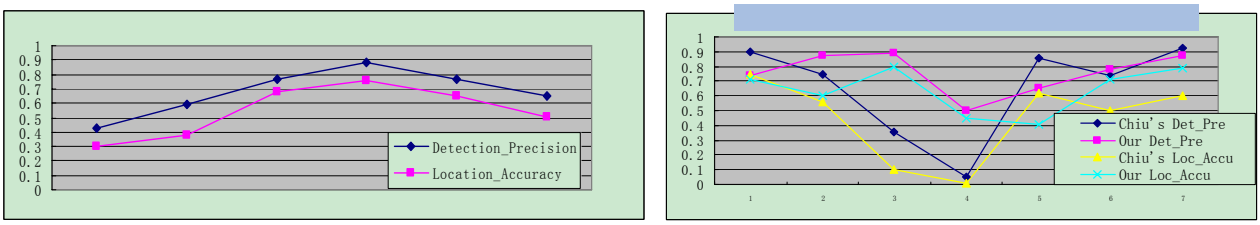
相关信息 / 视频集合	视频个数	分割出的亚镜头个数	播放时间长度	系统处理耗时	视频 MPG 文件占用的磁盘空间	特征文件占用的磁盘空间
源视频数据集	101	652500	58 小时	25 小时	100G 字节	227.69M 字节
查询视频数据集	30	25500	2 小时 20 分钟	56 分钟	5G 字节	7.22M 字节

如第 2 小节所述，适合拷贝检测的视觉特征必须满足鲁棒性，实时性和精简性要求。表 1 显示我们实验数据处理的具体情况，本文提出的拷贝检测系统在特征提取上能达到实时速度（即处理时间等于或者短于视频播放时间），并占用相对少量的磁盘空间，适合大规模的视频拷贝检测任务。

首先，我们使用 TRECVID 2008 拷贝检测比赛的评测指标[16]来验证提出的算法的鲁棒性。评测指标包括检测准确率（Det_Prec）和定位精度（Loc_Accu），其定义和计算方法如式（4）所示，

$$Det\_\Pr ec = \frac{\#Correct\_Det}{\#All\_Det} \qquad Loc\_Accu = \frac{\#Overlap(Det,Copy)}{\#Copy\_Frames} \tag{4}$$

其中#Overlap(Det,Copy)表示系统检测到的帧集合与真正的拷贝帧集合的交集，即检测正确的帧数目。在不同的 K 值下，我们测试得到不同的精度。如图 4（a）所示，在 K 等于 86 时对常见轻微拷贝变化取得将近 90%的平均检测准确率，表明我们提出的算法能够达到实用的要求。



(a) (b)

图 4（a）在各种轻微变化，不同 K 值下的平均检测性能。K 等于 86 时系统取得理想效果（b）在各种局部和时序拷贝变化下，我们的算法和[8]算法的指标曲线。我们的算法使用适合此数据集的最佳 K 值

Fig.4（a） Curves of average detection precision and location accuracy of our algorithm using different K (b) Curves of our algorithm and Chiu's [8] under various copy changes (average results). As in other BoW works, our K is tuned for best performance

其次，我们对比了本算法和其他算法的性能。Chiu 在文献[8]中提出的算法声称比先前的基于 OM 特征的研究[2][3]效果要好。这 3 项拷贝检测研究都是基于 OM 特征的。我们使用一些比较难处理的拷贝变化下的查询视频，分别使用本文提出的方法和[8]中的 DTW 算法进行检测，得到的评测指标曲线如图 4（b）所示。图 4（b）表明，不同于使用 OM 的方法，我们的算法使用局部稳定特征点的轨迹行为特征，能够对一些比较严重的局部变化（如遮挡，裁剪画面，画面平移和画中画等）鲁棒。这些局部图像变化改变了帧图像的部分内容，打乱了

OM 方法依赖的块序列关系，从而影响了 OM 方法的鲁棒性。我们的思路的本质是针对视频时空不变模式进行建模，把拷贝变化前后不变的视觉内容作为特征进行匹配，所以具有鲁棒性。但是，图 4（b）也表明我们的方法对一些时域变化并不能取得更好的效果，原因是当前使用的简单行为统计方法和 5 维直方图特征不能完全平滑去除时序变化的影响。

为了体现局部特征轨迹的效果，我们在图 5 中给出了在各种拷贝情况下亚镜头轨迹的形态，表明在拷贝变化后，轨迹行为的总体趋势相似。

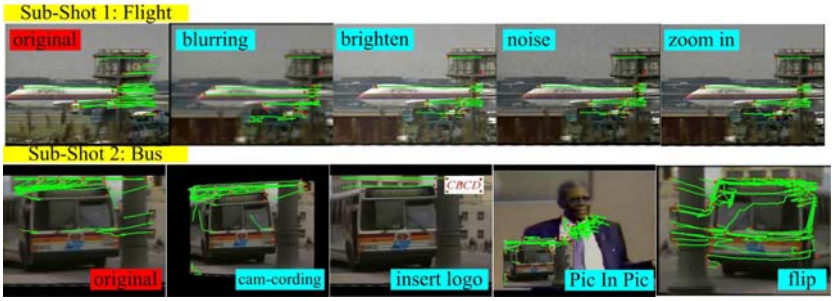


图 5 在各种常见拷贝变化下轨迹行为的一致性。一幅图像是一个亚镜头的代表帧。绿色曲线表示亚镜头的特征点轨迹

Fig.5 Trajectory behaviors are relatively similar despite of various usual copy transformations. Green curves represent trajectories

6. 总结和未来工作展望

如何提取鲁棒的视觉特征是基于内容的视频拷贝检测的一个关键研究点。本文的算法基于视频连续帧的时空轨迹行为，快速产生精简的不变模式特征，并使用轨迹视觉关键词技术，在亚镜头级别定位拷贝视频片段。标准数据集上的对比实验表明提出的算法对各种变化尤其是局部画面变化有较好的鲁棒性。以后的研究将专注于利用轨迹构造更加精细的特征应对时序的变化，以及提高轨迹特征的区分能力。

参考文献：

[1] A Joly, O Buisson, C Frelicot. Content-based Copy Retrieval using Distortion-based Probabilistic Similarity Search. IEEE Trans. on Multimedia, 2007.

[2] X. S. Hua, X. Chen, and H. J. Zhang, Robust video signature based on ordinal measure, ICIP, 2004.

[3] C. Kim and B. Vasudev, Spatiotemporal sequence matching for efficient video copy detection, IEEE Trans. on Circuits and Systems for Video Technology, 2005.

[4] K Mikolajczyk, T Tuytelaars, C Schmid, A Zisserman. A Comparison of Affine Region Detectors. IJCV, 2005.

[5] K Mikolajczyk, C Schmid. A Performance Evaluation of Local Descriptors. IEEE Trans. on PAMI, 2005.

[6] J Law-To, V Gouet-Brunet, O Buisson, N Boujemaa. Local Behaviours Labelling for Content Based Video Copy Detection. ICPR, 2006.

[7] Xiao Wu, AG Hauptmann, CW Ngo. Practical Elimination of Near-Duplicates from Web Video Search. MM, 2007.

[8] CY Chiu, CH Li, HA Wang, etc. A Time Warping Based Approach for Video Copy Detection. ICPR, 2006.

[9] A. Hampapur, K.-H. Hyun, and R. M. Bolle, Comparison of sequence matching techniques for video copy detection, The SPIE Conference on Storage and Retrieval for Media Databases, 2002.

[10] J Law-To, L Chen, A Joly, I Laptev, O Buisson. Video copy detection: a comparative study. CIVR, 2007.

[11] N Moenne-Loccoz, E Bruno, S Marchand-Maillet. Local Feature Trajectories for Efficient Event-Based Indexing of Video Sequences. LNCS, 2006.

[12] J Shi, C Tomasi. Good Features to Track. CVPR, 1994.

[13] M Takimoto, J Adachi. Scene duplicate detection from videos based on trajectories of feature points. MIR, 2007.

[14] J Sivic, A Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. ICCV, 2003.

[15] The origin of the video corpus is MUSCLE-VCD-2007 http://www-rocq.inria.fr/imedia/civr-bench/index.html

[16] Guidelines for the TRECVID 2008 CD task Evaluation. http://www-nlpir.nist.gov/projects/tv2008/tv2008.html