文章编号: 1006-2475(2016) 01-0054-05

基于轨迹预测的动态匿名算法

马佳仕,吕　鑫,戚荣志,张云飞,许国艳,刘　璇

(河海大学计算机与信息学院,江苏 南京 210098)

摘要: 位置 K 匿名是实现 LBS(Location Based Services) 隐私保护的重要手段。已有的 K 匿名机制大多针对无知识背景的攻击者模型,对攻击者能力的估计不足,存在用户位置隐私泄露的风险。针对此问题,本文提出一种基于历史轨迹预测的 LBS 动态匿名算法。该算法充分考虑攻击者基于历史数据对用户轨迹的预测能力,根据用户轨迹隐私泄露的风险级别,动态调整 K 匿名值实施保护,实验证明该算法在保护用户位置隐私方面是有效的。

关键词: 历史数据; 动态匿名算法; 轨迹预测

中图分类号: TP309.2　　文献标识码: A　　doi: 10.3969/j.issn.1006-2475.2016.01.012

Dynamic Anonymity Algorithm Based on Trajectory Prediction

MA Jia-shi, LYU Xin, QI Rong-zhi, ZHANG Yun-fei, XU Guo-yan, LIU Xuan

(College of Computer and Information, Hohai University, Nanjing 210098, China)

Abstract: K-anonymity for location privacy is an important solution to protect the user's trajectory in LBS. However, K-anonymity for location privacy can not always protect user's privacy if the attackers have any background knowledge. We propose a dynamic anonymity algorithm based on historical data trajectory prediction. This algorithm uses the trajectory prediction ability to define the background knowledge of attackers', when the user's trajectory has a predicted risk, changs the value of anonymity set. A series of experiments on synthetic datasets are made, the results show this algorithm is feasible and effective.

Key words: history data; dynamic anonymity algorithm; trajectory prediction

0 引　言

随着无线定位技术和移动设备的发展,位置服务[1] (LBS, Location Based Services) 已逐渐深入人们的日常生活。基于位置的服务是指服务提供商根据用户的位置信息提供各种服务,用户可主动获取路线导航、商业搜索,也可接受附近的商家提供的广告服务。然而,用户的位置和轨迹数据中含有大量的隐私内容,如用户的家庭工作地址、健康状况、生活习惯等。用户在享受 LBS 提供各类便捷服务的同时,也面临着极大的隐私泄露威胁。

一般来讲,位置服务中的隐私保护可以分为 2 种: 位置隐私和查询隐私[2]。如张某利用自己带有 GPS 的手机提出"寻找 5 min 内距离我最近的肿瘤医院"。这是导航系统中常见的查询服务。一方面,用户不想让任何人知道他现在所在位置(如"医院") ; 另一方面用户也不想让任何人获知自己提出了哪方面的查询请求,如与某特定肿瘤相关的医院查询。前者属于位置隐私保护范畴,后者属于查询隐私保护范畴。

通常,位置服务作为一种移动互联网应用服务为了满足用户的按需查询需要多次响应用户请求,根据查询的生命周期和实时性,可以分为快照空间查询和连续空间查询[3-4]。快照空间查询通常指用户"偶然"提交的空间查询,隐含每次查询请求中的标志符不能对应到同一实体。连续空间查询通常指用户在一段时间内连续地或周期性地提交查询请求,隐含每次查询请求都包括一致的用户标志。

为了解决位置服务中的隐私保护问题,位置 K

收稿日期: 2015-08-31

基金项目: 国家自然科学基金面上项目(61272543) ; 国家科技支撑计划项目(2013BAB06B04) ; 中国华能集团公司总部科技项目(HNKJ13-H17-04) ; 江苏省自然科学基金资助项目(BK20130852) ; 江苏省博士后科研资助计划项目(1401001C)

作者简介: 马佳仕(1990-) ,男,江苏盐城人,河海大学计算机与信息学院硕士研究生,研究方向: 网络信息安全; 吕鑫(1983-) ,男,博士后,研究方向: 密码学,网络信息安全; 戚荣志(1980-) ,男,讲师,博士,研究方向: 软件测试; 张云飞(1980-) ,男,讲师,博士,研究方向: 数据挖掘,知识工程; 许国艳(1971-) ,女,副教授,博士,研究方向: 大数据管理,数据起源追踪; 刘璇(1989-) ,女,博士研究生,研究方向: 数据挖掘,大数据安全。

匿名模型[5-7]是当前 LBS 系统中普遍采用的隐私保护模型: 当一个移动用户的位置无法与其他 K-1 个用户的位置相区别时,称此位置满足位置 K-匿名。实现位置匿名思想可以分为 3 种: 1) 发布假的位置信息,即不公布真实的位置信息[8]。2) 时空匿名技术,即降低对象位置的时空粒度,且用时空区域来表示对象真实的位置信息,其中匿名区域图形化的形状叫做匿名框。匿名框的大小与服务质量成反比,和匿名程度成正比。3) 基于数据加密技术,用数据加密技术来防止隐私泄露,能起到很好的效果,但是匿名时间相对较长,服务质量较差。潘晓等人[9]提出了隐私模型和质量模型,来确保 LBS 构建匿名框时有可靠地隐私保护能力和较好的服务质量。对于 LBS 服务中的查询隐私问题,主要的解决方法是查询 *l*-diversity 原则[10-13],这种模型要求对一个匿名区域提出的查询请求中,至少要包括 *l* 个不同的敏感值。以上各种研究都基于攻击者不具备背景知识[14-15]的前提,如果攻击者拥有大量历史数据,对这些数据推理分析,形成一定的背景知识,那么这些隐私保护模型可能不能很好地保护用户的隐私。

本文的工作主要包括: 1) 提出一种基于网格划分的轨迹预测方法; 2) 基于轨迹预测的动态匿名算法(DAA)的实现; 3) 通过实验得出算法是有效的。

1　基于网格划分的轨迹预测方法

现实中,一个人每天早上从家里出发,经过某些路径上班,然后又返回家中,人们一般性的活动轨迹总体来说是有规律的,这就为轨迹预测[16]提供可能。

通常移动设备收集的用户轨迹是包含时间戳的位置序列,它可以表示为 $T=\{ID,(t_1,x_1,y_1),(t_2,x_2,y_2),\ldots,(t_n,x_n,y_n)\}$,其中 ID 表示该轨迹的标识符,它代表移动对象、个体或某种服务的用户,(t_i,x_i,y_i) 表示移动对象在 t_i 时刻的位置为 (x_i,y_i)。由于移动设备能收集到的是用户在某个时刻的地理位置坐标信息,为了简化预测方法,可以采用空间网格划分方法。首先对包含所有轨迹采样点的二维空间进行划分[17],形成相同大小的许多单元格,对于大量的历史轨迹数据,每条轨迹都经过若干个单元格,每个单元格都维持一张单元格记录表如图 1 所示,记载该单元格经过的采样点信息,包括采样点所属的轨迹 ID、采样时间以及地理坐标。

定义 1　最长公共子序列。给定 2 个轨迹经过的单元格序列 S_1 和 S_2,如果存在子序列 S_{sub} 满足 $S_{sub}\subset S_1$,$S_{sub}\subset S_2$ 且不存在子序列 $S'_{sub}\supset S_{sub}$,满足上述条件,则称 S_{sub} 是 S_1 和 S_2 的最长公共子序列。

下面以一个例子阐述该预测方法: 对于某个正在使用位置服务的用户,假设当前该用户已经发起过多次服务请求,LBS 记录了该用户的轨迹信息 T_u(L1、L2、L7),那么对应于网格空间 3 个单元格 C1、C2、C7。遍历该网格空间,找出所有经过这些单元格的历史轨迹数据,形成候选历史轨迹集如图 2 所示。求出 T_u 与候选历史轨迹集的最长公共子序列,将含有最长公共子序列的历史轨迹集放入轨迹预测集内。可以发现,T_u 与候选历史轨迹集最大公共子序列就是 C1、C2、C7,那么得到包含这个最大公共子序列的轨迹预测集如图 3 所示。由于 C8 出现在 3 条历史轨迹中,而 C9、C10 分别只出现 1 次,所以该用户轨迹接下来有较大可能经过 C8,也可能经过 C9 或 C10。为了提高预测速度,可以选取某一个时期的历史数据作为预测参考,单元格范围也可以缩小至以 T_u 为中心的一定区域,在这个区域内寻找候选轨迹集,这种改进的策略是以降低预测精度为代价的,适用于对响应时间需求较高的场合。

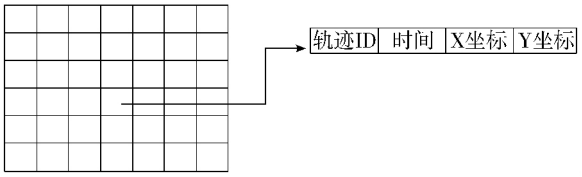

图 1　单元格记录表

T_1: C1、C2、C3、C4、C5
T_2: C1、C2、C3、C4
T_3: C1、C2、C3、C4
T_4: C1、C3、C4、C5
T_5: C1、C3、C4、C5
T_6: C1、C3、C4、C5、C6
T_7: C1、C2、C7、C8
T_8: C1、C2、C7、C8
T_9: C1、C2、C7、C8、C9
T_{10}: C1、C2、C7、C10
T_{11}: C1、C3、C4、C5

图 2　候选历史轨迹集

T_7: C1、C2、C7、C8
T_8: C1、C2、C7、C8
T_9: C1、C2、C7、C8、C9
T_{10}: C1、C2、C7、C10

图 3　轨迹预测集

2　基于轨迹预测的动态匿名算法

为了使 LBS 中的位置隐私保护问题得以解决,一般的方案是基于位置隐私保护的 K 匿名模型。当

攻击者不具备任何知识背景的情况下，能够识别某个用户对应的查询的概率为 1/K，但是当攻击者具备一定的知识背景，对被攻击者的一些历史情况有一定了解，那么攻击者能够识别某个用户对应的查询的概率就有可能大于 1/K。攻击者可以利用已有的知识，对被攻击者行为进行预测，如果发现被攻击者接下来的位置与预测值符合，那么攻击者就可能最终识别某个查询请求和某个用户的对应关系，导致用户的隐私泄露。

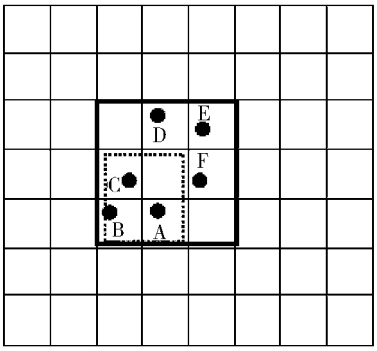

图 4 攻击示意图

如图 4 所示，实线代表匿名框，此时匿名集内有 6 个用户，假设攻击者想获取用户 A 的位置信息，如果攻击者没有预测能力，那么识别 A 的位置概率为 1/6，当攻击者成功推断出 A 可能在图中虚线框时，识别 A 的概率变为 1/3，因此 A 用户存在位置隐私泄露的风险。

对于上述问题，一个可行的方法就是 LBS 匿名服务器先于攻击者对用户的行为进行预测，匿名服务器拥有不低于攻击者背景知识的大量的历史轨迹数据，当匿名服务器认为该用户可能存在轨迹被提前预测的风险时，动态增加匿名集 K 的值，使识别某个用户对应的查询的概率更低，从而保护用户的隐私。潘晓等人认为，匿名服务器构建匿名集时，要满足隐私模型和质量模型，较大的 K 值可能带来匿名服务质量的降低，即需要花费更多的时间去寻找符合条件的用户请求参与匿名。因此当 LBS 匿名服务器认为用户轨迹不容易被预测时，还原默认的匿名集 K 值。

动态匿名算法(DAA，Dynamic Anonymity Algorithm)

输入：用户已形成的移动轨迹 T_u

输出：匿名集

1) 将用户轨迹 T_u 映射到空间网格，找到对应的单元格序列 C。

2) 寻找历史轨迹单元格序列与 C 的最大公共子序列 S_{sub}。

3) 若 S_{sub} 的长度与 T_u 长度的比值大于预测支持度阈值 β，则认为该用户的轨迹存在被预测的可能，表达式为：

$$p=\begin{cases}\text{true}, & |S_{sub}|/|T_u|\geq\beta\\ \text{false}, & |S_{sub}|/|T_u|<\beta\end{cases} \quad (1)$$

4) 当 p 为 true 时，对于用户的查询请求，采用较大的 K 构建匿名集，当 p 为 false，采用较小的 K 构建匿名集。

5) 返回此匿名集。

动态匿名算法流程如图 5 所示。

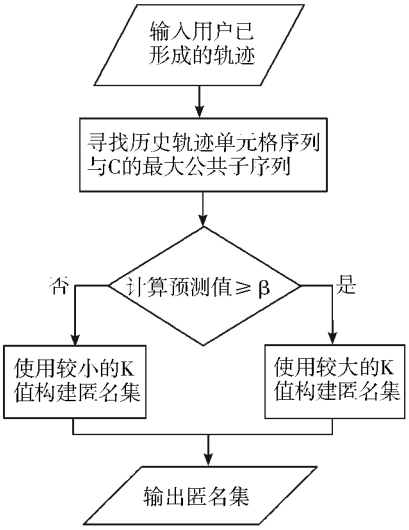

图 5 动态匿名算法流程

3 实验结果与分析

实验采用的数据集 OLDEN 由 Brinkhoff 基于网络的移动对象数据生成器[18] 生成，共包含100000条轨迹，生成的数据可以表示德国奥尔登堡(Oldenburg) 市的道路网络状况一天的移动数据(区域总面积 23.57 km × 26.92 km)。

实验环境：Intel(R) Core(TM) i5-5200U CPU @ 2.20 GHz；4 GB 内存；Windows 8 操作系统；算法是在 MyEclipse 环境下，基于 Java 语言编写。

实验方案：从匿名服务器对 LBS 请求的匿名集平均隐私保护强度、平均匿名时间方面验证本文提出算法的有效性。实验选取 D-TC 算法[19] 进行了对比，D-TC 算法采用抑制连续查询中容易暴露位置隐私的匿名集的构建来抵御查询跟踪攻击，二者都能抵御查询跟踪攻击，并且都是针对单个用户进行匿名，因而具有良好的可对比性。

首先将整个地图区域划分成 1000 × 1000 个单元格，选取数据集中 40000 条轨迹，共 2565797 个轨迹采样点作为历史轨迹数据。将所有历史轨迹数据映射到对应的单元格，另外选取 10000 条轨迹，共620494个数据点作为不同用户，每一个采样点模拟一次向 LBS 系统发起的查询。LBS 系统默认匿名集匿名参数 K = 5，即默认每个匿名集里包括 5 个不可区分的用户发起的查询请求。在本文提出的 DAA 算法中，若根据公式(1) 计算当前用户存在被预测可能

时,设增量 ΔK = 10,即匿名集匿名参数为 K + ΔK。

3.1　平均匿名集隐私保护强度

当攻击者无背景知识时,匿名集隐私保护强度 D = K。假设攻击者具备背景知识,在某个时刻能够预测用户的轨迹,那么能够识别某个用户对应的查询的概率大于 1/K,定义此时匿名集的隐私保护强度,表达式为:

$$D = k \cdot d, \quad d \in (0,1] \tag{2}$$

其中,d 代表匿名集对攻击者的防范能力,值在 0 到 1 之间。为了便于实验,假设 d = 0.5,即当攻击者能够利用历史轨迹信息预测用户下一个查询请求的大致位置时,匿名集的隐私防范能力设为 K/2。

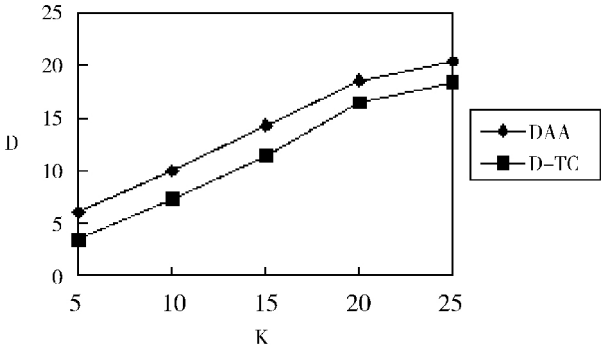

图 6　隐私保护强度

图 6 反映了 DAA 算法和 D-TC 算法在攻击者具备背景知识条件下,匿名集的平均隐私保护强度的对比情况。横坐标代表匿名集 K 的值,纵坐标代表匿名集平均隐私保护强度。

在相同 K 条件下,DAA 的匿名集隐私保护强度 D 都要大于 D-TC 算法,证明该算法在增强匿名集的平均隐私保护能力方面是有效的。对比不同 K 值的条件,2 种算法的匿名集隐私保护强度都随着 K 值增大而增大,并且 DAA 算法与 D-TC 算法渐渐逼近,说明在 K 值增大的情况下,DAA 算法的 ΔK 增量在匿名集 K + ΔK 值中的权重下降,对匿名集的隐私保护能力影响渐渐减弱,导致 2 种算法的性能趋于相似,表明 DAA 算法在 K 值较小的情况下提高匿名集平均保护强度的效果明显。

3.2　平均匿名所需时间

平均匿名时间是指平均完成一个查询请求构建匿名集所要消耗的时间。本组实验将 DAA 算法与 D-TC 算法进行对比,对于 D-TC 算法,时间代价主要表现在寻找 K 个用户查询请求所需要的时间,DAA 算法时间代价表现在轨迹预测和寻找 K 个匿名用户所需要的时间。如图 7 所示的实验结果表明,DAA 算法在匿名效率上略差于已有的 D-TC 算法,是因为 DAA 算法不但要完成寻找 K 个匿名用户的过程,还要额外完成轨迹预测的步骤,因此要花费更多时间。在 K 值较小的情况下,轨迹预测的时间代价明显,当 K 值较大时,寻找 K 个匿名用户所花费的时间占匿名时间总时间的比重上升,轨迹预测的步骤所需时间占匿名总时间的权重下降,因此在 K 值较大时,2 种算法性能趋于相似。

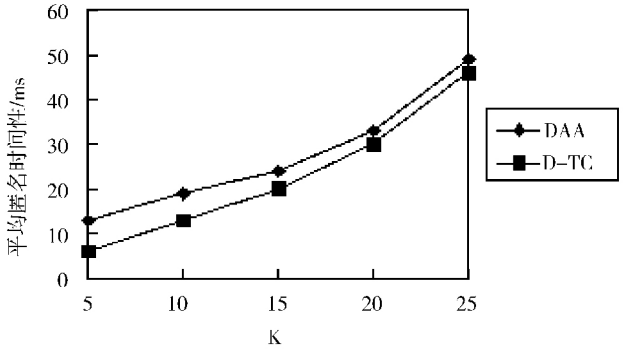

图 7　平均每个请求匿名时间

虽然 DAA 算法在匿名时间上效率并没有优于 D-TC 算法,但是之间的差异较小,也可以通过硬件设备加以解决,而用户更关心的不是平均匿名时间的微小差异,而是隐私是否被泄露。因此,综合上述实验,可以认为本文提出的基于历史轨迹预测的动态匿名算法 DAA 在保护用户位置隐私方面是有效的,平均匿名时间也是可接受的。

4　结束语

随着移动设备和定位技术的发展,产生大量移动对象轨迹数据。位置服务中,构建匿名框既要考虑匿名强度,也要满足较高的服务质量。本文提出的基于历史轨迹数据预测的轨迹隐私保护算法能够动态调整隐私保护强度,抵御查询跟踪攻击,保护位置隐私。从实验结果来看,算法能够达到预期的效果。然而目前的研究还有一定的不足,该算法旨在针对位置服务中出现的位置隐私问题,尚未解决可能存在的查询隐私问题,下一步的研究工作是进一步改进算法,在保护位置隐私的同时,防止查询隐私泄露。

参考文献:

[1] 霍峥,孟小峰. 轨迹隐私保护技术研究 [J]. 计算机学报, 2011,34(10):1820-1830.

[2] 潘晓,肖珍,孟小峰. 位置隐私研究综述 [J]. 计算机科学与探索, 2007(3):268-281.

[3] Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking [C] // International Conference on Mobile Systems, 2003. 2003:31-42.

[4] Xu Toby, Cai Ying. Location anonymity in continuous location based services [C] // Proceedings of the 15th ACM

International Symposium on Geographic Information Systems. 2007: Article No. 39.

[5] Sweeney L. K-anonymity: A model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002,10(5):557-570.

[6] Abul O, Bonchi F, Nanni M. Never walk alone: Uncertainty for anonymity in moving objects databases [C]// IEEE 24th International Conference on Data Engineering, 2008. 2008:376-385.

[7] Ghinita G, Kalnis P, Khoshgozaran A, et al. Private queries in location based services: Anonymizers are not necessary [C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008:121-132.

[8] Kido H, Yanagisawa Y, Satoh T. Protection of location privacy using dummies for location-based services [C]// IEEE 21th International Conference on Data Engineering Workshops. 2005:1248-1254.

[9] 潘晓,郝兴,孟小峰. 基于位置服务中的连续查询隐私保护研究 [J]. 计算机研究与发展, 2010,1(1):121-129.

[10] Xiao Zhen, Xu Jianliang, Meng Xiaofeng. p-Sensitivity: A semantic privacy-protection model for location-based services [C]// Proceedings of the 9th International Conference on Mobile Data Management Workshops. 2008:47-54.

[11] Zhou Bin, Pei Jian. The K-anonymity and *l*-diversity approaches for privacy preservation in social networks against neighborhood attacks [J]. Knowledge and Information Systems, 2011,28(1):47-77.

[12] Dondi R, Mauri G, Zoppis. The *l*-Diversity problem: Tractability and approximability [J]. Theoretical Computer Science, 2013,511(14):159-171.

[13] Tripathy B K, Maity A, Ranajit B, et al. A fast p-sensitive *l*-diversity anonymisation algorithm [C]// 2011 IEEE Recent Advances in Intelligent Computational Systems. 2011:741-744.

[14] 林欣,李善平,杨朝晖. LBS 中连续查询攻击算法及匿名性度量 [J]. 软件学报, 2009,20(4):1058-1068.

[15] 王彩梅,郭亚军,郭艳华. 位置服务中用户轨迹的隐私度量 [J]. 软件学报, 2012,23(2):352-360.

[16] 霍峥,孟小峰,黄毅. PrivateCheckIn: 一种移动社交网络中的轨迹隐私保护方法 [J]. 计算机学报, 2013,36(4):716-726.

[17] Gidofalvi G, Huang Xuegang, Pedersen T B. Privacy-preserving data mining on moving object trajectories [C]// 2007 International Conference on Mobile Data Management. 2007:60-68.

[18] Brinkhoff T. A framework for generating network-based moving objects [J]. Geoinformatica, 2002,6(2):153-180.

[19] Stenneth L, Yu P S. Global privacy and transportation mode homogeneity anonymization in location based mobile systems with continuous queries [C]// 2010 6th International Conference on Collaborative Computing: Networking, Applications and Worksharing. 2010:1-10.

(上接第 19 页)

[3] 任靖,李春平. 最小距离分类器的改进算法——加权最小距离分类器 [J]. 计算机应用, 2005,25(5):992-994.

[4] 刘相滨,邹北骥,孙家广. 基于边界跟踪的快速欧氏距离变换算法 [J]. 计算机学报, 2006,29(2):317-323.

[5] 郭磊,王秋光. Adaboost 人脸检测算法研究及 OpenCV 实现 [J]. 哈尔滨理工大学学报, 2009,14(5):123-126.

[6] 孙志. 基于 OpenCV 的人脸识别算法实验平台研究与实现 [D]. 长春: 吉林大学, 2014.

[7] 徐军. 基于混沌的灰度图像加密算法设计与分析 [D]. 长春: 吉林大学, 2012.

[8] Benedict R Gaster, Lee Howes, David Kaeli, 等. OpenCL 异构计算 [M]. 2 版. 张云泉,张先轶,龙国平,等译. 北京: 清华大学出版社, 2013.

[9] 陈钢,吴百锋. 面向 OpenCL 模型的 GPU 性能优化 [J]. 计算机辅助设计与图形学学报, 2011,23(4):571-581.

[10] 詹云,赵新灿,谭同德. 基于 OpenCL 的异构系统并行编程 [J]. 计算机工程与设计, 2012,33(11):4191-4195.

[11] HSA Foundation. HSA Programmer's Reference Manual [EB/OL]. http://www.hsafoundation.com/standards/, 2015-08-11.

[12] 吴兰. 基于 HSA 的 Kaveri 测试与优化 [D]. 苏州: 苏州大学, 2014.

[13] 赵成龙,施慧彬,俞忻峰. 基于 OpenCL 的双 GPU 基数排序算法 [J]. 计算机与现代化, 2015(1):27-30.

[14] 贾海鹏,张云泉,龙国平,等. 基于 OpenCL 的拉普拉斯图像增强算法优化研究 [J]. 计算机科学, 2012,39(5):271-277.

[15] Wang Guohui, Xiong Yingen, Yun Jay, et al. Accelerating computer vision algorithms using OpenCL framework on the mobile GPU-A case study [C]// 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013:2629-2633.

[16] 李焱,张云泉,王可,等. 异构平台上基于 OpenCL 的 FFT 实现与优化 [J]. 计算机科学, 2011,38(8):284-286,296.

[17] 周桐. 基于 PCA 的人脸识别系统的设计与实现 [D]. 哈尔滨: 哈尔滨工业大学, 2007.