JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 4, pp. 113-121, April 2020
https://doi.org/10.9708/jksci.2020.25.04.113

# Privacy-Preserving Method to Collect Health Data from Smartband

Su-Mee Moon*, Jong-Wook Kim*

*Student, Dept. of Computer Science, Sangmyung University, Seoul, Korea
*Professor, Dept. of Computer Science, Sangmyung University, Seoul, Korea

## [Abstract]

With the rapid development of information and communication technology (ICT), various sensors are being embedded in wearable devices. Consequently, these devices can continuously collect data including health data from individuals. The collected health data can be used not only for healthcare services but also for analyzing an individual's lifestyle by combining with other external data. This helps in making an individual's life more convenient and healthier. However, collecting health data may lead to privacy issues since the data is personal, and can reveal sensitive insights about the individual. Thus, in this paper, we present a method to collect an individual's health data from a smart band in a privacy-preserving manner. We leverage the local differential privacy to achieve our goal. Additionally, we propose a way to find feature points from health data. This allows for an effective trade-off between the degree of privacy and accuracy. We carry out experiments to demonstrate the effectiveness of our proposed approach and the results show that, with the proposed method, the error rate can be reduced upto 77%.

▸ **Key words**: Health Data Collection, Data Privacy, Local Differential Privacy

## [요 약]

센서 기술의 발전과 스마트 워치, 스마트 밴드와 같은 웨어러블 기기의 보편화로 개인의 건강 데이터를 실시간으로 수집하는 일이 가능해졌다. 웨어러블 기기에서 파생된 걸음 수, 심박 수와 같은 건강 데이터들은 모바일 환경의 위치, 날씨 데이터 등의 외부 데이터와 결합하여, 개인의 라이프 스타일 및 건강 상태를 분석하는 방식으로 활용되고 있다. 이처럼 웨어러블 기기에서 파생된 건강 데이터는 편리하고 유용한 기능을 제공하지만 개인의 생활과 밀접한 연관이 있기 때문에 외부에 노출될 경우 심각한 프라이버시 침해 문제가 발생한다. 이에 본 연구는 지역차분프라이버시와 특징점 추출 알고리즘을 사용하여, 웨어러블 기기에서 추출한 건강 데이터를 데이터 소유자의 프라이버시 침해 없이 데이터 수집가에게 전송할 수 있는 기법을 소개한다. 지역차분프라이버시를 통해 데이터 소유자의 프라이버시를 보호하였으며 특징점 알고리즘으로 프라이버시 보호 수준과 데이터 유용성간의 상충 관계를 조절하였다. 실험 결과는 제안하는 기법이 단순 방법에 비해 최대 77% 정도의 오차율 개선이 있음을 보여준다. 수집된 데이터는 데이터 사용자의 요구에 따라 헬스 케어 및 맞춤형 서비스 산업에서 유의미하게 활용될 수 있다.

▸ **주제어**: 건강 데이터 수집, 개인정보 보호, 지역 차분 프라이버시

• First Author: Su-Mee Moon, Corresponding Author: Jong-Wook Kim
*Su-Mee Moon (sumeedi@naver.com), Dept. of Computer Science, Sangmyung University
*Jong-Wook Kim (jkim@smu.ac.kr), Dept. of Computer Science, Sangmyung University
• Received: 2020. 02. 24, Revised: 2020. 04. 16, Accepted: 2020. 04. 16.


http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

## I. Introduction

현대 사회를 특징 짓는 연결성은 IoT 기술의 발전, 즉 센서의 다양성에서 발생한다. 센서는 활용 목적에 따라 각기 다른 방식으로 발전되어 왔다. 예를 들어 CCTV는 범죄를 예방하기 위한 목적에서, 스마트 밴드는 사용자의 운동량을 측정하기 위해 개발되었다. 스마트 밴드와 같은 웨어러블 기기 내부 센서는 사람이 착용할 수 있도록 소형으로 만들어진다. 또한 하나의 작은 기기에 가속도계, HRM(Heart Rate Monitoring) 등 다양한 센서가 내장되어 있다. 사용자가 웨어러블 기기를 착용하면 센서는 사용자의 상태를 감지하고 데이터를 저장한다. 이처럼 센서가 소형화, 다양화되면서 웨어러블 기기는 단순히 개인의 운동량을 측정하는 것이 아닌, 건강 데이터를 생산 및 전송하는 용도로 사용되고 있다.

웨어러블 기기에는 온도, 압력, 습도, 자외선과 같은 사용자 주변 환경뿐 아니라 사용자의 심박 수, 수면 상태, 걸음 수와 같은 건강 데이터를 측정하는 센서가 부착되어 있다. 예를 들어 스마트 밴드에는 가속도계가 내장되어 있기 때문에 걸음 수를 통해 사용자의 일상을 유추해낼 수 있다. Fig. 1은 걸음 수를 바탕으로 사용자가 집을 나선 시간, 사무실에서 근무하는 시간 등을 유추해낸 모습이다. 단순한 예로 정적인 상태는 특정 장소에 머무르고 있거나 교통수단으로 이동 중임을 뜻하며, 증가 상태는 도보로 이동 중 또는 운동하고 있음을 나타낸다. 이와 같이 정렬 및 정제 과정을 거친 데이터는 시그널 프로세싱 과정에서 머신러닝과 같은 방법으로 데이터 구조를 파악하고 분석하게 된다. 걸음 수 데이터 외에, 모바일 환경에서의 위치 데이터 및 사용자가 촬영한 사진 정보 등을 함께 사용하면 보다 정확한 예측이 가능하다. 이와 같이 걸음 수 데이터는 사용자의 생활 패턴을 파악할 수 있는 중요한 요소다.

사용자의 웨어러블 기기에 저장된 건강 데이터는 유용하게 사용하기 위해 데이터 수집자에게 전송된다. 데이터 수집자는 수집한 건강 데이터를 서비스 개발자에게 전송한다. 서비스 개발자는 머신러닝 알고리즘을 기반으로 데이터를 가공하여 사용자에게 맞춤형 서비스를 제공한다. 심박 수, 걸음 수와 같은 건강 데이터는 맞춤형 헬스케어 어플리케이션으로 이용될 수 있다. 예를 들어 만성폐쇄성 폐질환 환자의 신체 활동을 통해 라이프 스타일을 파악하면, 환자의 신체 행동이 건강에 미치는 영향을 알 수 있다 [1]. 치매 역시 일상적인 신체 활동의 영향을 받는 질병 중 하나이므로 신체 활동 분석을 통해 적절한 치료를 내릴 수 있다 [2]. 건강 데이터를 위치, 날씨 데이터 등의 외부 데이터와 함께 가공하면 사용자의 생활 패턴을 분석하여 모

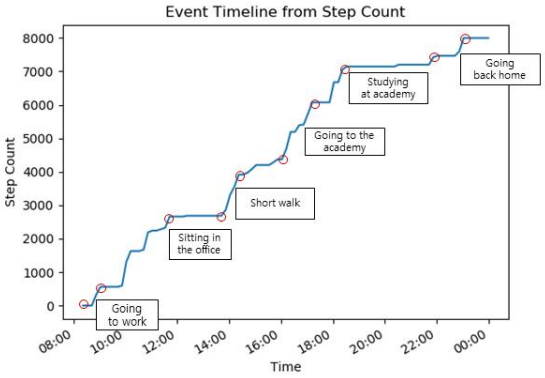



Fig. 1. Event Timeline from Step Count

바일 컨시어지 서비스로 활용 가능하다. 또한 가속도계를 이용하면 노인 또는 거동이 불편한 사람의 응급 상황을 감지할 수 있다. 즉 건강 데이터는 서비스 제공자에게 서비스 개발을 위한 정보를, 사용자에게는 편리함을 제공한다.

웨어러블 기기에서 파생되는 건강 데이터는 유용한 반면 개인의 사생활과 밀접하게 연관되어 있기 때문에 외부에 유출될 경우 심각한 사생활 침해가 발생한다. 2018년 7월 싱가포르에서는 150만 명의 의료 데이터와 16만 명의 의약품 처방 정보가 탈취되었으며, 중국 암시장에서는 의료 관련 데이터베이스가 2천 달러 미만에 판매되고 있다. 웨어러블 기기를 통해 추출되는 건강 데이터도 마찬가지로 유출의 위험에 노출되어 있다. 앞서 유출되어 오·남용된 의료 데이터의 사례처럼 심박 수, 수면 상태, 걸음 수 등의 건강 데이터 또한 유출되어 사생활 문제로 이어질 가능성이 높다. 이처럼 데이터 유출에 대한 사회적 이슈로 인해 현대 사회에는 정보 제공에 대한 부정적 인식이 팽배해있다. 이로 인해 서비스 제공자는 대량의 데이터를 습득할 수 없으며 사용자에게 고품질의 서비스를 제공할 수 없다. 따라서 사용자의 사생활을 보호하면서 건강 데이터를 활용할 수 있는 방법에 대한 논의가 필수적으로 요구된다.

웨어러블 기기를 통해 안전하게 사용자 데이터를 전송하기 위한 방법으로, 지역 차분 프라이버시 기법이 존재한다. 지역 차분 프라이버시는 데이터 수집에서 사용자의 정보를 보호하기 위해 사용되는 최신의 접근법이다. 지역 차분 프라이버시에 의해 사용자는 직접 차분 프라이버시를 만족하도록 데이터를 변조하여 서비스 제공자에게 전송한다. 사용자 입장에서는 변조된 데이터를 전송하므로 건강 데이터 유출의 위험이 사라지며, 서비스 제공자 입장에서는 원본 데이터를 전송받지 못하므로, 유출에 대한 부담이 사라짐과 동시에 대량의 데이터를 전송 받아 고품질의 서비스를 제공할 수 있게 된다.

본 연구는 데이터 소유자(제공자)의 프라이버시를 보존하면서 효과적으로 데이터를 수집하는 방법에 대해 제안한다. 건강 데이터를 소유자 측면에서 노이즈를 추가하여 변조한 후, 변조된 데이터를 데이터 수집자에게 전송하는 방식이다. 제안하는 방법을 통해 데이터 소유자의 프라이버시를 보존하면서 유의미한 통계 데이터를 활용할 수 있다. 본 연구의 구성은 다음과 같다. 첫 번째 관련 연구에서는 건강 데이터 활용 방법 및 선행 연구를 설명한다. 두 번째 배경 및 문제 정의에서는 차분 프라이버시 및 지역 차분 프라이버시 기법을 소개한 후 실험에서 사용한 데이터와 해결하고자 하는 문제를 정의한다. 세 번째 알고리즘에서는 소유자가 건강 데이터에서 특징점을 추출하여 수집자에게 전송하는 방법에 대해 언급한다. 네 번째 본문은 지역 차분 프라이버시와 특징점 탐색 알고리즘을 사용하여, 스마트밴드로 수집한 걸음 수 데이터를 소유자 프라이버시를 보호하면서 수집하는 과정에 대해 설명하고 실험 결과를 보인다. 마지막으로 결론에서는 연구 결과와 차후 연구에 대해 언급할 것이다.

## II. Related Work

웨어러블 기기에서 추출할 수 있는 건강 데이터는 심박수, 심전도, 혈압, 산소 포화도, 수면 상태, 걸음 수, 칼로리 등으로 기기의 종류에 따라 다양하다. 대표적인 건강 데이터의 활용으로 지능형 어플리케이션과 사용자 행동 인식 연구가 존재한다. 지능형 어플리케이션은 사용자가 웨어러블 기기를 통해 건강 상태, 운동 경과를 모니터링하거나 일상을 디지털 공간에 저장하는 라이프 로깅 어플리케이션을 의미한다. CodeBlue [3]는 심박수, 심전도, 산소 포화도, 맥박을 모니터링하는 방법을 제안한다. 그리고 AlarmNet [4]과 Medical MoteCare [5]은 혈압과 산소 포화도 같은 생리적 정보와 온도와 빛 등의 환경 요소를 함께 발전시키는 연구를 진행했다. 사용자 행동 인식 영역은 건강 데이터를 통해 사용자의 움직임을 인식하여 사용자 상태를 파악하는 것이다. 예를 들어 응급 상황 알림 시스템은 거동이 불편한 장애인 및 노인을 대상으로 장시간 움직임이 없거나 비정상적인 움직임을 감지했을 때 자동으로 외부에 도움을 요청해준다. 이와 같은 시스템은 웨어러블 기기에 내장 되어 있는 관성 센서 데이터 분석을 통해 이루어진다. 사무실 환경에서 일상적인 행동 정보를 분석하는 일도 사용자 행동 인식 영역에 포함된다.

건강 상태 모니터링은 전통적으로 활발히 연구되어 온 분야이다. 건강 데이터를 통해 맞춤 의료 서비스와 조기 질병 진단이 가능하다. 예를 들어 병원에서 환자의 건강 데이터를 수집하면, 의사는 다른 환자들의 진료 기록 외에 환자 개인의 현재 상태를 기반으로 판단을 내릴 수 있게 되므로 보다 정확한 판단이 가능해진다 [6]. 최근에는 심전도 측정으로 심장질환 상태를 파악하거나 심박 수 변화를 통해 사용자가 겪는 스트레스 정도를 예측한다. Taelman et al. [7]는 스트레스 정도와 사용자의 심박수 변화 사이의 연관관계를 조사했다. 루벤 카톨릭 대학교에서 28명의 학생을 대상으로 한 실험으로, 평온한 사진에 비해 멘사 시험을 학생에게 제시했을 때 평균 심박 수는 73.52에서 75.94까지 상승했다. 이는 심적 스트레스 상황이 심장 리듬을 자극한다는 것을 의미한다. Fisher et al. [8]은 울혈성심부전을 갖고 있는 환자의 심박 수를 웨어러블 기기를 통해 추적하여, 원격으로 모니터링하는 방법에 대해 소개한다. 가속도계로 신체 활동을 측정하여 운동량과 건강 상태를 동시에 파악하는 방법도 연구되고 있다. 여러 연구들을 통해 건강과 운동에는 상당한 관계가 있음이 밝혀졌으며, 여가 시간에 더 활동적인 사람이 사망 확률이 낮았다 [9], [10].

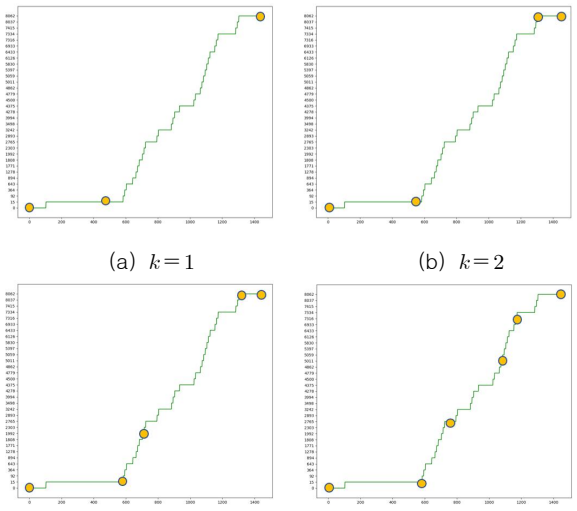
건강 데이터는 건강관리를 목적으로 한 모니터링뿐 아니라, 머신러닝 등의 방법으로 건강 데이터를 분석하여 신체 활동을 분류하는 인간 행동 인식 영역에서도 활발히 연구되고 있다. 예를 들어 Altun et al. [11]은 소형 관성 센서와 자기 센서로 19가지의 인간 행동을 분류하는 연구를 진행했다. 건강 데이터를 이용한 행동 분류는 주로 환자나 노인을 위한 서비스로 활용된다. Melillo et al. [12]는 심전도를 통해 저혈압에 의한 쓰러짐 등을 감지하고 예방하는 방법에 대해 소개한다.

## III. Background and Problem Statement

### 1. Local Differential Privacy

차분 프라이버시는 어떤 유형의 배경지식에도 익명성이 보장되는 프라이버시 보호 모델이다. 차분 프라이버시에서 소유자는 믿을 수 있는 제 3자인 데이터 수집자에게 원본 데이터를 보내고, 데이터 수집자는 차분 프라이버시를 만족하도록 원본 데이터에 노이즈를 추가한다.

반면 지역 차분 프라이버시에서는 믿을 수 있는 제 3자가 존재하지 않는다고 가정한다 [13], [14]. 소유자는 직접 지역 차분 프라이버시를 만족하도록 노이즈를 원본 데이터에 추가하여 변조된 데이터를 데이터 수집자에게 전송한



(a) $k=1$ (b) $k=2$ (c) $k=3$ (d) $k=4$

Fig. 2. An example of feature points searched by proposed method for different $k$

다. 임의화 알고리즘 $A$는 $v_1, v_2 \in Dom(A)$이고 $A$의 가능한 모든 출력이 $O$일 때 $\varepsilon$-지역 차분 프라이버시를 만족한다. 이는 다음과 같은 수식으로 나타낼 수 있다 [15], [16].

$$\Pr[A(v_1)=O] \le e^{\varepsilon} \times \Pr[A(v_2)=O]$$

변조된 데이터 $O$를 받은 수집자는 수집자가 미리 갖추고 있던 배경 지식과는 상관없이 원본 데이터가 $v_1$ 혹은 $v_2$인지 확신할 수 없다. 여기서 $\varepsilon(\varepsilon>0)$은 프라이버시와 유용성 사이의 트레이드 오프 관계를 형성한다. $\varepsilon$이 높으면 노이즈를 적게 추가하기 때문에 프라이버시 수준이 낮지만 데이터 활용 측면에서 유용성이 높다. 반면에 $\varepsilon$이 낮으면 노이즈가 많이 추가되기 때문에 프라이버시 수준이 높지만 유용성이 낮다. $\varepsilon$-지역 차분 프라이버시에서는 소유자가 직접 노이즈를 추가하여 수집자에게 보내므로, 소유자간 서로 다른 $\varepsilon$을 설정함으로써 개인화된 프라이버시 수준 설정이 가능하다 [17]. 또한 한 명의 소유자가 여러 개의 포인트를 수집자에게 보내는 상황에서 전송하는 포인트의 개수를 $n$이라고 할 때, 소유자는 원본 데이터에 $\varepsilon/n$을 적용한 노이즈를 추가하여 변조된 데이터를 보내게 된다. (Fig. 3)

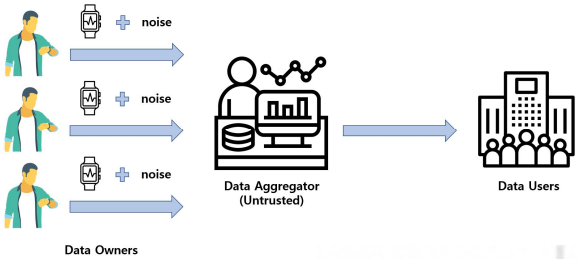



Fig. 3. Local differential privacy

## 2. Problem Statement and Straightforward Solution

본 연구는 데이터 소유자의 프라이버시를 보호하면서 Fig. 1과 같은 단조 증가 형태의 건강 데이터를 데이터 수집자에게 효과적으로 전송하는 방법에 대해 제안한다. 스마트 밴드를 착용한 데이터 소유자를 $U=\{u_1,u_2,\cdots,u_w\}$라고 할 때, $w$는 총 소유자 수다. 소유자 $i$의 특정 하루 동안 누적 걸음 수 데이터를 $s_i=((t_1,x_1),(t_2,x_2),\cdots,(t_n,x_n))$로 표현하고 $t_n$은 시간을, $x_n$은 걸음 수를 의미한다. $n$은 미리 정의된 시간 범위를 따르며 $[x_{\min},x_{\max}]$의 데이터를 사용한다. $s_i$에 지역 차분 프라이버시를 만족하도록 노이즈를 추가한 결과를 $s'_i=((t_1,x'_1),(t_2,x'_2),\cdots,(t_n,x'_n))$라고 할 때, 노이즈가 추가된 $x'_n$은 라플라스 매커니즘에서 $\varepsilon$을 $\varepsilon/n$로 설정하여 도출된 값이다. 이는 다음과 같은 수식으로 표현할 수 있다.

$$x'_d = x_d + Lap\left(\frac{\Delta s}{\varepsilon/n}\right)$$

수식에서 전역민감도인 $\Delta s$는 $x_{\max}-x_{\min}$을 의미한다.

데이터 소유자가 지역 차분 프라이버시를 만족시키면서 자신의 건강 데이터를 수집자에게 보내는 여러 방법이 존재한다. 가장 쉬운 방법인 단순 기법에서 데이터 소유자는 데이터 수집자에게 $s_i$에 노이즈를 추가한 $s'_i$를 전송하게 된다. $s'_i$에서 $x'_d$는 $\varepsilon$을 $n$개로 나누어 각각의 $x_d$에 $\varepsilon/n$을 적용한 노이즈를 더한 값이다. 데이터 수집자는 $U$로부터 $w$만큼의 $s'_i$를 수집하게 된다. $S=\{s'_1,s'_2,\cdots,s'_w\}$라고 할 때, 각각의 $s'_i$의 $t_d$에 따른 $x'_d$의 평균인 $AVG$의 $x_d$는 다음과 같은 수식으로 나타낼 수 있다.

$$AVG(x_d)=\frac{1}{w}\times\sum_{s'_i\in S}x'_d$$

# IV. Proposed Method to Collect Health Data from Smartband Users

건강 데이터는 종류에 따라 1시간에 한 번 측정되기도 하지만, 1분 간격으로 측정되기도 한다. 예를 들어 심박수를 1분 간격으로 측정될 경우 하루에 얻을 수 있는 데이터양은 1,440개이다. 만약 단순 기법을 통해 건강 데이터를 보낸다면, $x_i$에 추가되는 노이즈는 $\varepsilon$을 1,440개로 나

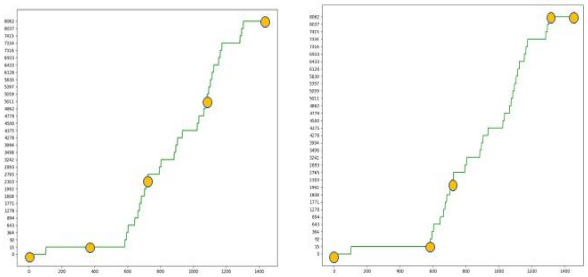
누어 적용한 값이다. 차분 프라이버시에서 $\varepsilon$이 낮을수록 노이즈를 많이 추가하기 때문에 유용성 또한 낮아지게 된다. 이처럼 3.2장에서 언급한 단순 기법을 이용할 경우 $\varepsilon$이 낮아지게 되어 건강 데이터 유용성이 현저히 떨어진다. 이에 소유자 프라이버시를 보호하면서 데이터의 유용성을 취하기 위해서는 모든 데이터가 아닌 적절한 개수의 특징점을 보내는 방법을 사용해야 한다. 본 장에서는 프라이버시 보호 건강데이터 수집 방법을 데이터 소유자 측면과 데이터 수집자 측면으로 분류하여 설명한다.

### 1. Data Owner's Device-side Processing

데이터의 특징점을 수집하는 방법은 건강 데이터의 종류에 따라 다르게 나타난다. 예를 들어 심박 수와 같이 상승 및 하강하는 건강 데이터는 기울기 변화 지점을 이용해 특징점을 탐색할 수 있다. 연속하는 세 점의 기울기의 부호 차이를 통해 심박 수 데이터의 변화 지점을 탐색할 수 있고, 이를 특징점으로 하여 데이터 수집자에 전송하는 방법이다 [18]. 반면에 걸음 수와 같은 단조 증가 형태의 건강 데이터는 기울기 부호의 변화 지점이 없기 때문에 심박 수와 다른 방법으로 특징점을 탐색하여 데이터 수집자에게 전송해야 한다. $k$개의 특징점을 탐색하기 위한 방법으로는, $x$축을 $k-1$개로 나누어 도출해내는 기법과 $k$개의 특징점을 그룹화하여 탐색하는 기법이 존재한다. 전자가 단순 계산 방법이라면 후자가 본 연구에서 제안하는 기법이다. (Fig. 4) 단순 계산 기법은 시간이 적게 소요되기는 하지만 데이터의 성격을 반영하지 못하기 때문에 정확도와 활용 측면에서 효율적이지 않다. 반면 제안하는 기법은 데이터의 성격을 반영하며 특징점을 탐색하므로 단순 계산 방법에 비해 데이터 흐름을 보다 정확하게 유추해낼 수 있다.

제안하는 기법에서는 데이터 소유자의 디바이스가 자신의 데이터에서 $k$개의 특징점을 탐색한다. 예를 들어 Fig. 2에 마크되어 있는 지점들이 $k$를 1, 2, 3, 4로 설정했을 때의 특징점 탐색 결과다. 만약 $k$가 2라면, 알고리즘을 통해 두 점을 탐색하고 첫 번째 점과 마지막 점을 포함시켜 총 4개의 특징점을 반환한다. 따라서 $k$가 1, 2, 3, 4일 때, 실제로 반환되는 특징점의 개수는 3, 4, 5, 6이다.

$i$번째 소유자의 건강 데이터 시퀀스는 $s_i = ((t_1,x_1),(t_2,x_2),\ldots,(t_n,x_n))$이며 여기서 $n$은 하루 동안 측정한 걸음 수의 개수다. 데이터 소유자의 디바이스가 특징점 탐색 알고리즘을 사용하여 $s_i$의 일부를 데이터 수집자에게 전송한다고 할 때, $k$개의 특징점은 $P_i = ((t_a,x_a),(t_b,tx_b),\ldots,(t_k,x_k))$로 나타낸다. 특징점 알고리즘의 의사코드인 Fig. 6을 보면, 먼저 특징점을 구하기 위한 임시 리스트 $P_{list}$와 $k$개의 포인트 중 마지막에서 두 번째 포인트인 $slp$, $k$개 포인트의 마지막 탐색 지점인 $EndPoint_{list}$을 구한다 (line 1). 예를 들어 $n$이 1440이고 $k$가 3이면 $EndPoint = \{1437,1438,1439\}$이다. 그리고 $P_{list}$에서 마지막 포인트인 $P[slp+1]$를 $EndPoint[slp+1]$까지 한 점씩 이동시키면서 $error_{\min}$와 $C_{list}$를 업데이트한다 (lines 6 – 12). $P[slp+1]$를 마지막 점까지 이동시키면 $P[slp]$를 한 점 이동시키는 방식이다 (line 13). 이 때 $P_{list}$를 $slp$부터 검사했을 때, 만약 $P[h]$가 $EndPoint[h]$와 같다면, 끝까지 탐색했다는 의미이므로 $P[h-1]$을 한 점 이동시키고 $P[h]$를 $P[h-1]$ 다음 점으로 이동시킨다 (lines 14 – 19). 이 과정은 $P[0]$이 $EndPoint[0]$과 같을 때, 즉 $P_{list}$가 모든 경우를 탐색할 때까지 반복한다 (line 5). 탐색이 종료되면 특징점을 저장한 리스트인 $C_{list}$를 반환한다 (line 21). 데이터 소유자는 도출된 특징점에 노이즈를 추가한 $P'_i = ((t_a,x'_a),(t_b,x'_b),\ldots,(t_k,x'_k))$를 데이터 수집자에게 하루에 한 번 전송한다. 이때 노이즈는 라플라스 매커니즘에서 $\varepsilon$을 $\varepsilon/k$로, $\Delta s$를 $x_{\max}-x_{\min}$로 설정하여 추출한다.



(a) (b)

Fig. 4. (a) Naive approach (b) Proposed approach

Fig. 5은 데이터 소유자 측면의 과정을 보여준다. 먼저 특징점을 추출하고 특징점에 노이즈를 추가하여 수집자에게 전송한다. 이 때 소유자는 $k(k<n)$개의 특징점만을 수집자에게 전송하므로 전체 데이터 $n$을 전송했을 때보다 노이즈를 적게 추가한 변조 데이터를 보내게 된다. 즉 제안한 기법을 통해 데이터 수집자는 데이터 소유자의 사생활을 보호하면서 건강 데이터를 활용할 수 있다.

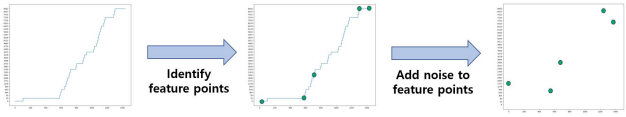



Fig. 5. Owner-side process of proposed approach

**Algorithm 1** Searching Feature Points

**input:** The number of feature points $k$, Sequence of a health data $s_i$
**output:** Feature points $C_{list}$

$P_{list} \leftarrow NULL$
$slp \leftarrow GetSecondtoLastPoint(k)$
$Endpoints_{list} \leftarrow GetEndPoints(k)$
$error_{min} \leftarrow \infty$
**while** $P[0] < Endpoints[0]$ **do**
  **for** $h \leftarrow P[slp]+1$ to $ListSize(s_i)$ **do**
    $curerror \leftarrow GetCurError(P_{list})$
    **if** $curerror < error_{min}$ **then**
      $error_{min} \leftarrow curerror$
      $UpdatePoints(C_{list}, P_{list})$
    **end if**
  **end for**
  $P_{slp} \leftarrow P_{slp}+1$
  **for** $h \leftarrow slp$ to 0 **do**
    **if** $P[h] = Endpoints[h]$ **then**
      $P[h-1] \leftarrow P[h-1]+1$
      $P[h] \leftarrow P[h-1]+1$
    **end if**
  **end for**
**end while**
**return** $C_{list}$

Fig. 6. Pseudo-code for searching feature points with step count data

### 2. Data Aggregator-side Processing

데이터 소유자가 특징점을 전송하면 데이터 수집자는 특징점을 잇는 직선을 구한다. Fig. 7은 데이터 수집자 측면의 과정을 나타내는 구조도이다. 우선 여러 소유자로부터 받은 건강 데이터의 특징점을 직선화한다. 만약 전송받은 특징점이 5개라면, 총 4개의 직선을 구하여 연결한다. 그리고 각각의 $s'_i$의 $t_d$에 따른 $x'_d$의 평균인 $AVG_{est}$의 $x_d$를 구한다. 즉 데이터 수집자는 $AVG_{est} = \{(t_1,x_1),(t_2,x_2),\ldots,(t_n,x_n)\}$를 얻게 된다. 여기서 $AVG_{est}$의 $x_d$는 데이터를 전송한 소유자의 수가 $w$라고 할 때, 각 점에 대한 $x'_d$를 $d$가 0일 때부터 $w$까지 더한 후 $w$로 나눈 평균을 의미한다. 각 점에 대해 도출된 통계 결과는 데이터 사용자의 요청에 따라 배포된다.

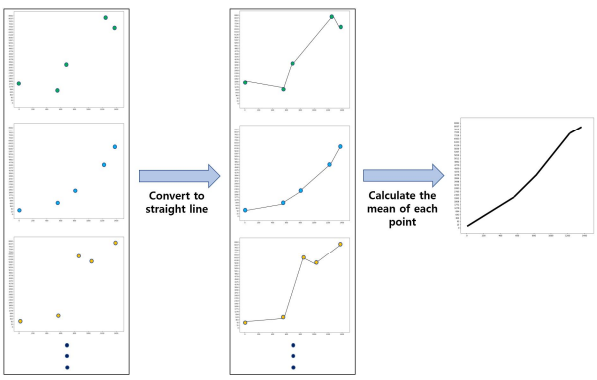



Fig. 7. Aggregator-side process of proposed approach

## V. Experiments and Results

### 1. Experiments

본 연구에서 제안하는 프라이버시 보존 건강 데이터 수집 방법을 평가하기 위해, 건강 데이터이자 단조 증가 형태인 누적 걸음 수 데이터로 실험을 진행했다. 실험에 사용된 데이터는 샤오미 미밴드를 통해 290명으로부터 10시에서 21시 사이에 수집한 1일 누적 걸음 수로, 데이터 사이즈에 따른 성능 평가를 위해 10, 100배 크기로 복제하여 사용했다. 또한 $\Delta s$는 6,000으로 설정하고 $\varepsilon$은 $\varepsilon=0.5$, $\varepsilon=1.0$, $\varepsilon=2.0$인 세 가지 경우에 대해 실험했다. 본 연구에서 알고리즘 성능을 평가하기 위해 정의한 오차율 $e$는 다음과 같다.

$$e = \frac{1}{n} \times \sum_{d=1}^{n} \left| AVG_{actual}(x_d) - AVG_{est}(x_d) \right|$$

$e$는 데이터 사이즈만큼 원본 데이터의 평균과 변조된 데이터의 평균의 차를 더한 후, 데이터 사이즈로 값을 나눈 결과이다. 오차율이 적을수록 원본 데이터와 변조된 데이터 사이의 차이가 적다는 뜻이므로 통계 결과가 유용함을 의미한다. Table 1은 특징점 4개를 탐색했을 때, $\varepsilon$과 데이터 사이즈에 따른 오차율을 나타낸다. $\varepsilon$이 높을수록 원본 데이터에 추가되는 노이즈가 적다는 것을 의미하므로 $\varepsilon=2.0$일 때 오차율이 가장 낮으며 $\varepsilon=0.5$일 때 $e$가 가장 높다. 또한 데이터 사이즈가 $290\times10^2$일 때, $290\times10^1$인 경우 보다 오차율이 낮다. 이를 통해 $\varepsilon$값과 데이터 사이즈가 클수록 오차율이 낮다는 것을 확인할 수 있다. 하지만 $\varepsilon$값과 프라이버시 보호 정도는 트레이드 오프 관계를 이루고 있으므로, $\varepsilon$이 높을 때 데이터 유용성은 높지만 프라이버시 보호 정도는 낮다. 4장에서 언급한 특징점 알고리즘인 단순 계산 기법과 본 논문에서 제안하는 기법을 해당 오차율을 통해 비교하고자 한다.

Table. 1. An example of average error for proposed approach ($k=4$)

| $\varepsilon$ / data size | 0.5 | 1.0 | 2.0 |
|---|---|---|---|
| $290\times10^1$ | 297.8282 | 147 | 76.78243 |
| $290\times10^2$ | 13.72048 | 9.739918 | 5.835464 |

### 2. Results

Fig. 8은 데이터 소유자 측면에서 단순 계산 기법과 제안하는 기법으로 특징점을 추출한 후, 데이터 수집자 측면 프로세싱을 거쳐 도출해 낸 $AVG_{est}$를 그래프로 나타낸 예

시이다. 제안하는 기법의 그래프가 단순 계산 기법의 그래프보다 원본 그래프와 유사하다는 것을 확인할 수 있다.

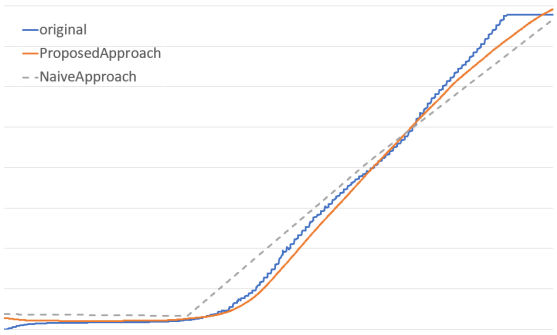
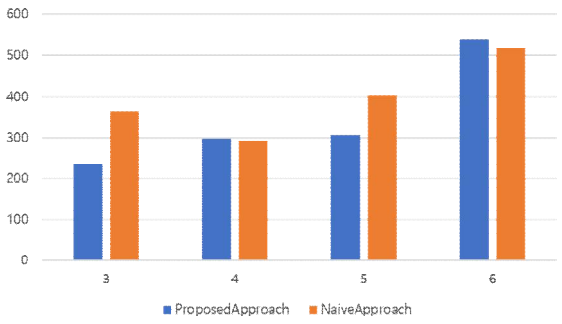


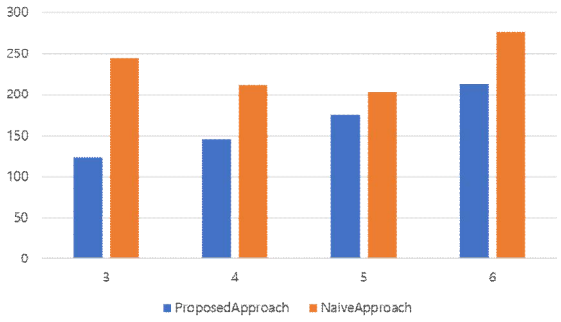

Fig. 8. An example of average error comparison between proposed and naive approach (data size = $290\times10^1$, $\varepsilon=2.0$, $k=4$)

Fig. 9는 데이터 사이즈를 고정시키고 $\varepsilon$값을 다양하게 설정하였을 때, 제안하는 기법과 단순 계산 기법을 비교한 막대그래프다. x축은 특징점 k의 개수이고 y축은 오차율을 나타낸다. 그래프를 통해 $\varepsilon$값이 커질수록 두 기법의 오차율이 줄어들며 $\varepsilon$이 1.0 이상일 때, 단순 계산 기법보다 제안하는 기법의 오차율이 적다는 것을 확인할 수 있다.

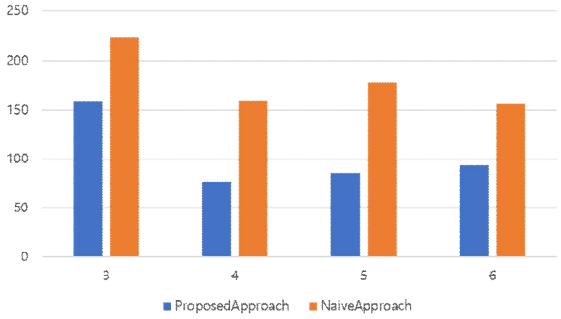
Fig. 10은 $\varepsilon$값을 고정시켰을 때, 데이터 사이즈에 따른 오차율의 변화를 보여준다. 데이터 사이즈가 10배 증가했을 때 오차율도 약 10배 줄어드는 것을 확인할 수 있다. 또한 제안하는 기법이 단순 계산 기법보다 오차율이 적은데, 이는 제안하는 기법이 원본 데이터와 보다 유사하며 데이터 유용성 측면에서 우수하다는 것을 의미한다.




(a) $\varepsilon=0.5$

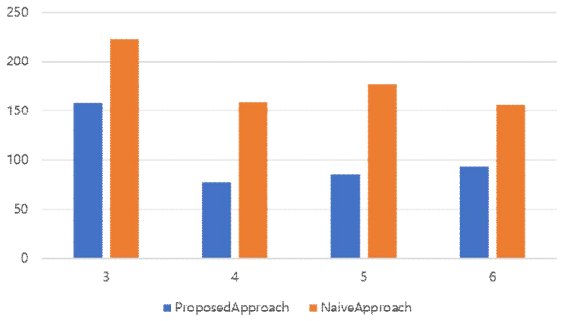



(b) $\varepsilon=1.0$

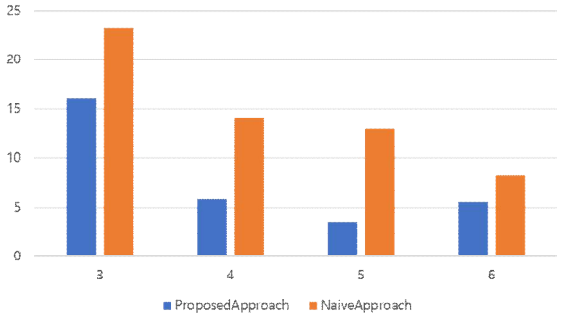



(c) $\varepsilon=2.0$

Fig. 9. Average error comparison between proposed and naive approach for various privacy budget (data size = $290\times10^1$)




(a) data size = $290\times10^1$




(b) data size = $290\times10^2$

Fig. 10. Average error comparison between proposed and naive approach for different data size ($\varepsilon=2.0$)

## VI. Conclusions and Future Work

스마트 밴드를 통해 수집될 수 있는 심박 수, 걸음 수, 수면 상태 등의 건강 데이터는 건강 상태 모니터링, 사용자 행동 인식, 컨시어지 서비스 등 다양한 영역에서 활용될 수 있는 유용한 데이터이지만 제 3자에게 오·남용 될 수 있는 민감한 데이터다. 따라서 건강 데이터의 유용성을 확보하면서 소유자 프라이버시를 보호하기 위한 방법이 필요하다. 본 연구에서는 데이터 소유자의 스마트 밴드에서 특징점을 추출하고 지역 차분 프라이버시를 만족하도록 노이즈를 추가하여 데이터 수집자에게 전송하는 프로세스를 통해 데이터 소유자의 프라이버시를 보호하면서 건강 데이터를 수집하는 방법을 제안하였다. 이 과정에서 특징점을 추출하기 위한 기법 두 가지를 살펴봤으며, 5장에서 본 논문이 제안하는 기법이 단순 계산 기법보다 성능이 우수하다는 것을 오차율을 통해 증명하였다. 차후 연구에서는 제안하는 특징점 추출 기법을 발전시켜, 해당 데이터에 적합한 특징점 개수를 탐색한 후 데이터 수집자에게 전송하는 방법을 통해 오차율을 낮추고 유용성을 향상시키는 방향으로 나아가고자 한다.

## REFERENCES


[1] R. Benzo, "Activity monitoring in chronic obstructive pulmonary disease," Journal of cardiopulmonary rehabilitation and prevention, vol. 29, p. 341, 2009.

[2] B. Chikhaoui, B. Ye, and A. Mihailidis, "Ensemble Learning-Based Algorithms for Aggressive and Agitated Behavior Recognition," in Ubiquitous Computing and Ambient Intelligence: 10th International Conference, UCAmI 2016, San Bartolomé de Tirajana, Gran Canaria, Spain, November 29–December 2, 2016, Part II, 2016, pp. 9-20.

[3] K. Missen, J. E. Porter, A. Raymond, K. de Vent, and J.- A. Larkins, “Adult deterioration detection system (adds): An evaluation of the impact on met and code blue activations in a regional healthcare service,” Collegian, 2017.

[4] M. M. M. Fouad, N. El-Bendary, R. A. Ramadan, and A. E. Hassanien, “Wireless sensor networks: a medical perspective,” Wireless Sensor Networks: From Theory to Applications, 2013.

[5] K. F. Navarro, E. Lawrence, and B. Lim, “Medical mote- care: A distributed personal healthcare monitoring sys- tem,” in eHealth, Telemedicine, and Social Medicine, 2009. eTELEMED’09. International Conference on. IEEE, 2009, pp. 25–30.

[6] Manogaran, G., Varatharajan, R., Lopez, D., Kumar, P. M., Sundarasekar, R., & Thota, C. (2018). A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system. Future Generation Computer Systems, 82, 375-387.

[7] J. Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel. Influence of mental stress on heart rate and heart rate variability. In J. Vander Sloten, P. Verdonck, M. Nyssen, and J. Haueisen, editors, 4th European Conference of the International Federation for Medical and Biological Engineering, pages 1366–1369, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[8] Fisher, R., Smailagic, A., & Sokos, G. (2017, December). Monitoring Health Changes in Congestive Heart Failure Patients Using Wearables and Clinical Data. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1061-1064). IEEE.

[9] Warburton, D. E., & Bredin, S. S. (2019). Health Benefits of Physical Activity: A Strengths-Based Approach.

[10] Ruegsegger, G. N., & Booth, F. W. (2018). Health benefits of exercise. Cold Spring Harbor perspectives in medicine, 8(7), a029694.

[11] Altun, K., Barshan, B., & Tunçel, O. (2010). Comparative study on classifying human activities with miniature inertial and magnetic sensors. Pattern Recognition, 43(10), 3605-3620.

[12] P. Melillo, R. Castaldo, G. Sannino, A. Orrico, G. de Pietro, and L. Pecchia, “Wearable technology and ecg processing for fall risk assessment, prevention and detection,” in Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 7740–7743.

[13] Ye, Q., Hu, H., Meng, X., & Zheng, H. (2019, May). PrivKV: Key-value data collection with local differential privacy. In 2019 IEEE Symposium on Security and Privacy (SP) (pp. 317-331). IEEE.

[14] Lim, J. H., & Kim, J. W. (2019). Privacy-Preserving IoT Data Collection in Fog-Cloud Computing Environment. Journal of the Korea Society of Computer and Information, 24(9), 43-49.

[15] Zhang, Z., Wang, T., Li, N., He, S., & Chen, J. (2018, January). Calm: Consistent adaptive local marginal for marginal release under local differential privacy. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (pp. 212-229).

[16] Kim, J. W., Kim, D. H., & Jang, B. (2018). Application of local differential privacy to collection of indoor positioning data. IEEE Access, 6, 4276-4286.

[17] Kim, J. W., Lim, J. H., Moon, S. M., & Jang, B. (2019). Collecting Health Lifelog Data From Smartwatch Users in a Privacy-Preserving Manner. IEEE Transactions on Consumer Electronics, 65(3), 369-378.

[18] Kim, J. W., Jang, B., & Yoo, H. (2018). Privacy-preserving aggregation of personal health data streams. PloS one, 13(11).

## Authors

Su-Mee Moon received the B.S. degree from Sangmyung University in 2019, where she is currently pursuing the master's degree with the Department of Computer Science. Her research mainly focuses on data privacy and Artificial Intelligence.

Jong-Wook Kim received the Ph.D. degree in Computer Science Department, Arizona State University, in 2009. He was a Software Engineer with the Query Optimization Group, Teradata, from 2010 to 2013. Dr. Kim is currently an Associate Professor with the Department of Computer Science at Sangmyung University. His primary research interests include the area of data privacy, distributed databases, and query optimization.