사용자 재실 여부에 기반한 사용자 태스크 단위의 데이터 스트림 분할 기법

김태준 ○, 손희석, 이동만
전산학부, 한국과학기술원
xowns5979@kaist.ac.kr, heesuk.son@kaist.ac.kr, dlee@kaist.ac.kr

A Task-level Data Stream Segmentation Method based on User Presence in a Smart Space

Taejun Kim○, Heesuk Son, Dongman Lee
School of Computing, KAIST

요 약

스마트 공간 속 사용자 태스크를 정확하게 인지하기 위해서는 태스크 단위의 센서 데이터 스트림을 안정적으로 확보하는 것이 중요하다. 대부분의 기존 연구들은 이를 위해 머신러닝 기법을 활용하지만, 이는 학습에 비용이 많이 든다는 점과 학습 된 모델이 다른 환경에 재사용되기 힘들다는 단점을 가지고 있다. 이를 극복하기 위해 본 연구에서는 스마트 공간 속 사용자 재실 유무 정보만을 활용하여 태스크 단위 데이터 스트림 분할을 수행하고 그 정확도를 보정하기 위해 분할된 데이터 스트림 사이의 유사성 비교를 통해 잘못 분할 된 데이터 스트림을 병합하는 기법을 제안한다. 제안하는 시스템의 성능 평가를 위해서 세미나실 테스트베드에서 수집된 데이터를 활용하여 데이터 스트림 분할 정확도를 측정하였다.

1. 서 론

스마트 홈, 스마트 시티 등 스마트 공간에서의 데이터 수집과 그에 기반한 맞춤형 서비스 제공은 산업 및 학계에서 많은 관심을 받고 있다. 이를 위한 핵심 기술로써 사용자의 행동 패턴을 분석하는 Group Activity Recognition에 대한 연구가 활발히 진행되고 있는데, 이들 중 상당수가 사용자 태스크 (Group Activity) 단위로 분할 된 스트림 데이터의 패턴을 분석한 후 테스트 데이터 스트림이 들어오면 이를 특정 사용자 태스크로 인지하는 방식을 따른다. 이 때 사용자 태스크 모델 및 머신러닝 알고리즘 설계가 인식 결과에 많은 영향을 미치지만, 그에 못지 않게 중요한 요소가 공간에서 수집되는 데이터 스트림을 사용자 태스크 단위로 정확히 분할하여 안정적인 데이터 셋을 확보하는 것이다.

Hidden Markov Model(HMM) 은 이러한 목적으로 사용되는 대표적인 머신 러닝 모델이다. San-Segundo[1]는 HMM 분석에 기반하여 일상의 활동에서 수집되는 데이터를 분할하는 방법을 제시하였고, Kabir[2]는 low-level 센서 데이터와 high-level 태스크 간의 연결(Mapping)을 표현해주는 Two-layer HMM을 제안하였으며, Saeedi[3]는 분할(Segmentation)에서 나타날 수 있는 여러 문제들을 간단하고 효율적으로 해결하는 segmented iHMM (siHMM) 을 제안하였다. 하지만 한편으로 머신러닝 기법은 충분한 양의 안정적인 데이터 셋을 확보하는 데에 비용이 많이들 뿐 아니라, 외부 관찰자를 통한 데이터 태깅이 사생활 침해 문제를 야기할 수도 있다. 또한, 데이터 수집 공간의 도메인이나 설치 된 센서 구성이 바뀔 때마다 기존에 학습된 모델을 재사용하지 못하고 데이터 수집부터 모델 학습까지 다시 수행해야 한다는 단점이 있다[4]. 따라서 이러한 단점들을 극복하는, 새로운 환경에서도 재사용 될 수 있으며 가벼운 (Light-weight) 데이터 스트림 분할 기법이 제시되어야 한다.

본 논문에서는 센서 데이터 스트림을 공간 별 빅데이터 셋의 학습 없이 사용자 재실 유무 정보만을 활용하여 사용자 태스크 단위로 분할, 저장하는 경량화 된 시스템을 제안한다. 시스템은 크게 두 개의 프로세스로 구성되는데, 첫 번째 프로세스에서는 수집된 공간 데이터 스트림을 사용자 재실 여부를 기준으로 분할한다. 이는 실제 공간에서 이뤄지는 대부분의 태스크가 사용자가 공간에 있을 때 수행되기 때문에 사용자가 공간에 존재하는 순간을 캡처하기 위함이다. 그러나 재실 여부에 따라 센서 데이터를 분할하더라도 사용자의 일시적 부재상태 (예: 화장실, 휴식 등) 로 인해 잘못된 데이터 분할을 수행할 가능성이 존재한다. 이러한 이슈를 보완하기 위해 두 번째 프로세스에서는 일시적 부재로 인한 데이터 분할을 발견 및 상쇄함으로써 데이터 분할 정확도를 보정하는 작업을 수행한다. 제안하는 시스템의 성능을 검증하기 위해 세미나실에서 수집된 8일 동안의 센서 스트림 데이터를 분할하여 그 정확도를 측정하는 실험을 수행하였다.

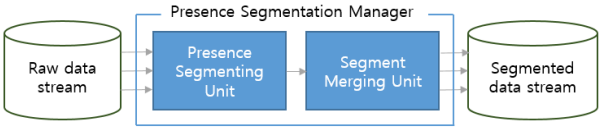

그림 1. Presence-based Segmentation 설계 개요

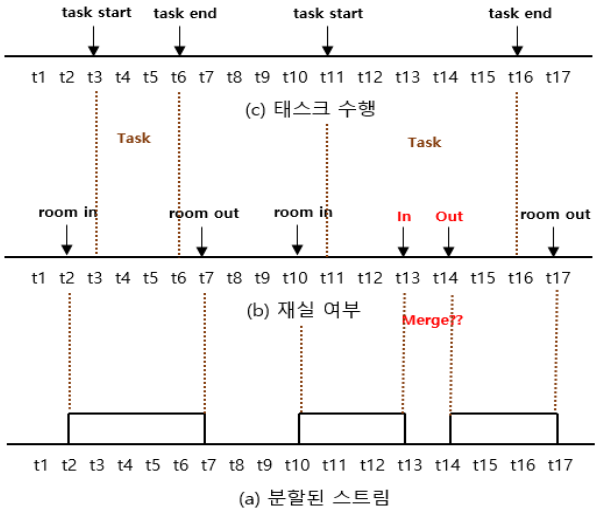

그림 2. 데이터 스트림 분할 및 병합 예제

정확도 분석 결과, Precision 0.838, Recall 0.904, Accuracy 0.770의 정확도가 측정되었다.

2. 논의

본 논문에서 실험에 활용한 공간은 세미나실로, 공부, 미팅, 토론, 식사, 휴식, 발표와 같이 사용자가 분명한 목적을 가지고 방문하여 활용하는 특수목적용 공간으로 볼 수 있다. 휴게실, 강의실, 침실과 같은 공간 또한 이러한 특징을 가지는 공간으로 간주된다. 하지만 제안된 시스템이 이러한 특수목적용 공간이 아닌, 사람들이 공간에 들어와도 별다른 태스크를 수행하지 않을 가능성이 충분한 공간 (예: 야외 열린 공간, 이동 통로 등) 에 적용 될 경우, 태스크가 아님에도 분할이 이루어지는 경우가 다수 발생하여 그 정확도가 매우 떨어질 가능성이 존재한다. 이를 보완하기 위해서는 의미가 없는 사용자 행위를 감지할 수 있는 추가적인 센서의 설치 및 데이터 해석 알고리즘 등이 추가로 도입되어야 한다.

3. 시스템 설계

그림 1은 본 논문에서 제안하는 시스템의 설계 개요이다. 첫 번째 컴포넌트인 Presence Segmenting Unit에서는 전달된 Raw data stream 에서 공간에 사용자가 없다가 누군가 입실을 한 경우에 분할을 시작하고 공간에 있는 모든 사람이 나갈 때 분할을 끝낸다. 이러한 작업을 반복적으로 수행하면 그림2와 같이 공간에 사용자가 존재하는 기간을 따라 분리 된 다수의 데이터 스트림이 생성된다. 이 때 재실 유무 정보를 제공해주는 센서는 공간의 물리적 구성에 따라 상이할 수 있기 때문에, 각 공간 별로 물리적 특성에 맞게 재실 여부를 판별할 수 있는 데이터 처리 모듈 (예: 도어 센서, 재실 센서 등) 을 제공해주어야 한다.

두 번째 컴포넌트인 Segment Merging Unit은 재실 유무에 기반하여 분할 된 데이터 스트림 중에서 동일한 태스크 수행 도중 일시적 사용자 부재로 인해 잘못된 분할이 이뤄진 경우를 발견하고 이러한 공백 (일시적 부재 기간) 전 후의 데이터 스트림을 다시 병합(Merge) 시켜주는 작업을 수행한다. 이를 위해, 우선 공백이 충분히 짧은 경우에 대하여 이러한 병합이 필요한지의 여부를 판단하게 된다. 이 때 '충분히 짧은 값'의 기준은 공간에서 주로 일어나는 태스크와 사용자 부재 상황의 유형에 따라 알맞은 값이 적용된다. 하지만 공공재인 세미나실이 연속적으로 예약 되어 서로 다른 태스크를 수행하는 선예약자와 후예약자의 교체 간격이 짧은 경우는 이렇듯 단순한 시간계산 만으로는 정확한 병합을 수행할 수 없다. 이런 단점을 극복하기 위해 Segment Merging Unit은 공백 전 후에 위치한 데이터 스트림 조각이 포함하는 상황 정보의 유사성을 비교하여 두 데이터 스트림이 같은 태스크의 연속적인 수행을 나타내는 것으로 판단되면 두 데이터 스트림의 병합을 수행한다. 이 때 유사성을 판단하는 수단으로 다차원 공간 속 포인트들의 거리 계산에 널리 활용되는 Euclidean distance를 이용한다. 먼저 각 스트림이 포함하는 각 센서 데이터 (예: 사운드레벨, 사용자 수 등) 를 대표할 수 있는 값 (예: 데이터 스트림 기간 동안의 평균값) 을 계산하여 공백 전후의 데이터 스트림을 표현해주는 벡터 포인트를 생성한 후 두 포인트 사이의 Euclidean distance를 계산한다. 만약 계산 결과가 기정의된 Threshold 값보다 작을 때 둘을 충분히 유사하다고 판단하여 같은 태스크로의 병합을 수행한다. 이 때 활용되는 Threshold 값은 실험에 의해 발견되는 최적값을 활용한다.

4. 시스템 구현 및 실험

4.1 구현 및 실험 환경

제안하는 시스템의 검증을 위해 본 실험에서는 표1에 정리 된 센서들이 설치 된 분산 미들웨어 [5] 기반 세미나실 테스트베드에서 8일간 사용자들의 공간 사용을 관찰하면서 수집되는 센서 데이터 스트림을 저장하였다. Segment Merging Unit이 공백의 합병 여부 판단을 위해 비교하는 간격 기준으로는 5분을 설정하였다. 즉, 공백이 5분 이하일 때에 유사성 판단을 수행한다. 공백 전후 스트림의 상황 정보 유사성 계산에는 각 스트림이 포함하는 사운드 레벨의 평균 값과 동시에 앉아있는 최대 의자 수가 이용되었다. 데이터 스트림의 합병 기준이 되는 Threshold 값은 가장 높은 결과를 반환해주는 2.5로 설정하였다.

4.2 정확도 분석

정확도 분석에는 이진 분류 기법(binary classification)을 사용하였으며, 이에 활용 된 변수들의 정의는 아래와 같다:

- True Positive (TP): 분할되어야 하는 시점에서 (태스크가 시작 또는 종료되는 시점)에서 분할이 정상적으로 수행한 경우
- True Negative (TN): 분할하면 안되는 시점 (태스크가 시작 또는 종료되는 시점이 아닌 곳) 에서 분할을 수행하지 않은 경우
- False Positive (FP): 분할하면 안되는 시점에서 분할을 수행한 경우
- False Negative (FN): 분할되어야 하는 시점에서 분할을

표 1. Precision, Recall and Accuracy

	TruePositive	FalsePositive	FalseNegative
Day1	16	4	2
Day2	16	4	0
Day3	8	4	0
Day4	12	4	2
Day5	18	0	4
Day6	18	4	2
Day7	24	2	2
Day8	2	0	0
합	114	22	12

Precision : 0.838 Recall : 0.904 Accuracy : 0.770

수행하지 않은 경우

이때, 시간 축 상에서 태스크를 수행한 구간을 제외하면 거의 대부분의 구간이 태스크를 수행하지 않은 구간이기 때문에, TN는 정확도 계산에 포함시키지 않는다. 위의 변수를 토대로 Precision, Recall, Accuracy를 계산하여 본 논문에서 제안하는 시스템의 정확도를 평가하였다.

그림3은 제안하는 시스템의 정확도 평가 결과를 보여준다. 실험 결과에 기록 된 대부분의 FalsePositive는 대부분 사용자가 무의미하게 공간을 사용할 경우 발생한다. 예를 들어, 사용자가 아무 의미 없이 잠시 공간에 들어왔다가 나가는 경우 아무런 태스크를 수행하지 않았기 때문에 데이터 스트림을 분할하는 것이 무의미하다. 하지만 제안하는 시스템은 사용자의 재실여부를 기준으로 분할을 수행하기 때문에 이를 태스크 데이터 스트림으로 분할한 것이다. FalseNegative는 대부분 재실 상태가 계속 유지되는 동안 동일한 사용자 그룹 (또는 개인) 이 하나 이상의 태스크를 연속적으로 수행할 때 (예: 간단한 식사를 하고 이어서 회의를 하는 경우) 발생하였다. 즉, 상이한 태스크를 진행하였기 때문에 데이터 스트림 또한 분할되어야 하지만 사용자 재실 여부만으로는 이를 정확히 판별할 수 없는 것이다.

그림 4는 분할된 데이터 스트림을 상황 유사도 계산을 토대로 병합해야 하는 케이스들이 발견 된 사례를 정리해 둔 표이다. 이에 따르면 8일간의 데이터 중에 병합 여부를 판단해야 하는 5분 이하의 공백은 Day3, Day5, Day6에 각각 한번씩 나타났다. 모두 공백 전후의 스트림이 같은 태스크를 수행하여서 병합이 수행되어야 하는 경우인데, 모두 전 후의 Euclidean Distance가 Threshold (2.5) 보다 작게 나와 병합이 수행되다. 전체 데이터 셋에서 많은 케이스가 발견되지는 않았지만 발생한 경우에는 모두 올바르게 동작하였다.

표 2. Segment Merging Cases

	Time	Max of TotalSeat	SoundAverage
Day3	15:03:05~15:52:28	1	49.86156
	15:54:00~16:22:01	1	47.60705
Euclidean distance : 2.2			
Day5	12:02:01~12:04:14	2	51.84024
	12:04:47~12:28:31	2	52.05204
Euclidean distance : 0.2			
Day6	17:24:09~17:26:48	0	50.61199
	17:27:25~17:28:05	0	49.28131
Euclidean distance : 1.4			

5. 결론

본 논문에서는 사용자의 재실 유무 정보만을 활용해서 공간 속 센서 데이터 스트림을 사용자 태스크 단위로 분할하여 저장하는 경량화 된 (Light-Weight) 데이터 분할 시스템을 제안하였다. 이에 더하여, 재실 유무 만으로 데이터 스트림을 분할할 경우 발생할 수 있는 오류를 보정하기 위한 목적으로 데이터 스트림 사이의 공백 공간 및 각 스트림의 상황적 유사도를 활용하여 동일한 태스크에 대한 데이터 스트림을 병합하는 프로세스를 제안하였다. 제안하는 시스템의 성능 평가를 위해 수행 된 실험에서는 세미나실 테스트베드 공간에서 수집된 8일간의 데이터 셋을 통해 0.838의 Precision, 0.904의 Recall, 0.770의 Accuracy 값을 측정하였다. 뿐만 아니라, 8일 동안 발생했던 상황 유사도 기반 데이터 스트림 병합 케이스들 또한 모두 올바르게 처리됨을 확인하였다. 본 연구의 확장 방안 (Future work)으로는, 실험을 통해 발견 된 False positive와 False negative의 원인을 해결하기 위해 의미없는 사용자 태스크를 식별하고 동일한 사용자 그룹의 사용자 태스크 전환을 분할할 수 있는 기법을 추가적으로 고안할 계획이다.

6. 사사문구

본 연구는 정부(미래창조과학부)의 재원으로 정보통신기술연구진흥센터의 지원을 받아 수행된 연구임 [2017-0-00537, 공간지능을 위한 IoT 사물간 자율협업 기술 개발]

참고문헌

[1] San-Segundo, Rubén, et al. "Segmenting human activities based on HMMs using smartphone inertial sensors." *Pervasive and Mobile Computing* 30 (2016): 84-96.

[2] Kabir, M. Humayun, et al. "Two-layer hidden markov model for human activity recognition in home environments." *International Journal of Distributed Sensor Networks* (2016).

[3] Saeedi, Ardavan, et al. "The segmented ihmm: A simple, efficient hierarchical infinite hmm." *arXiv preprint arXiv:1602.06349* (2016).

[4] Riboni, Daniele, et al. "Unsupervised recognition of interleaved activities of daily living through ontological and probabilistic reasoning." *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016.

[5] Son, Heesuk, et al. "A distributed middleware for a smart home with autonomous appliances." *Computer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual*. Vol. 2. IEEE, 2015.