라이프로깅 영상 스트림의 서술 생성을 이용한 자동 사건 분절

Automatic Event Segmentation based on Descriptions Generated from Life-Logging Image Stream

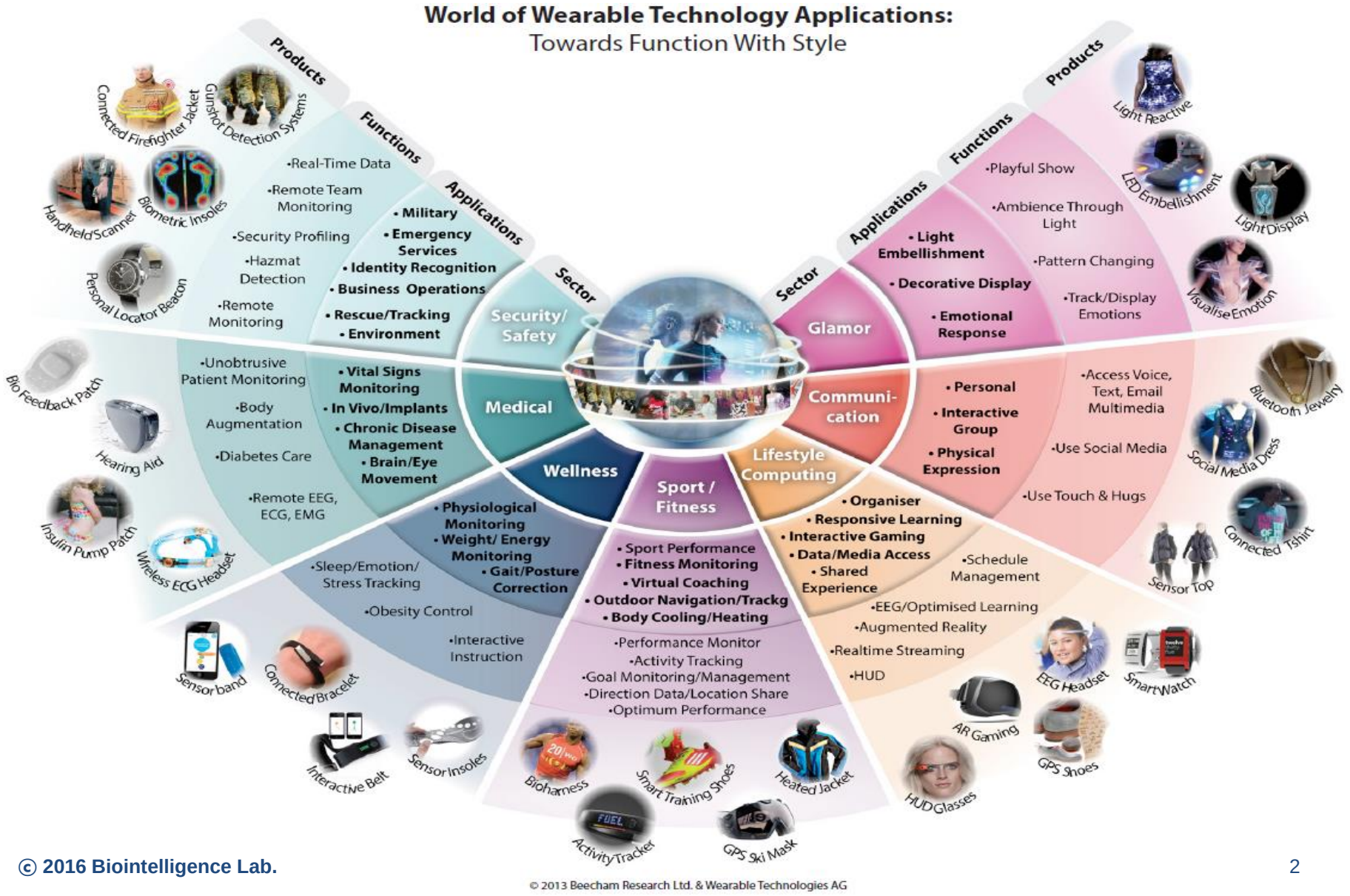
이충연, 한동식, 장병탁

한국컴퓨터종합학술대회(KCC)

2016. 7. 1

서울대학교 컴퓨터공학부 바이오지능연구실

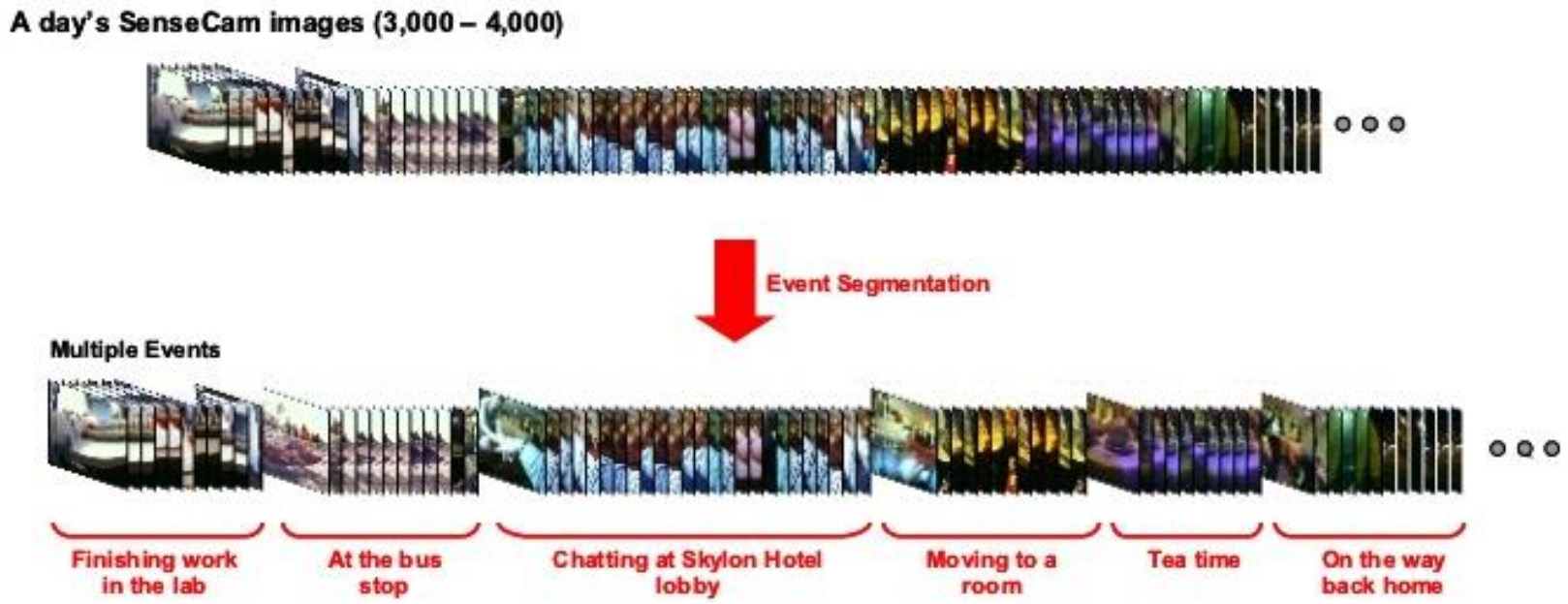
Motivation and Goals



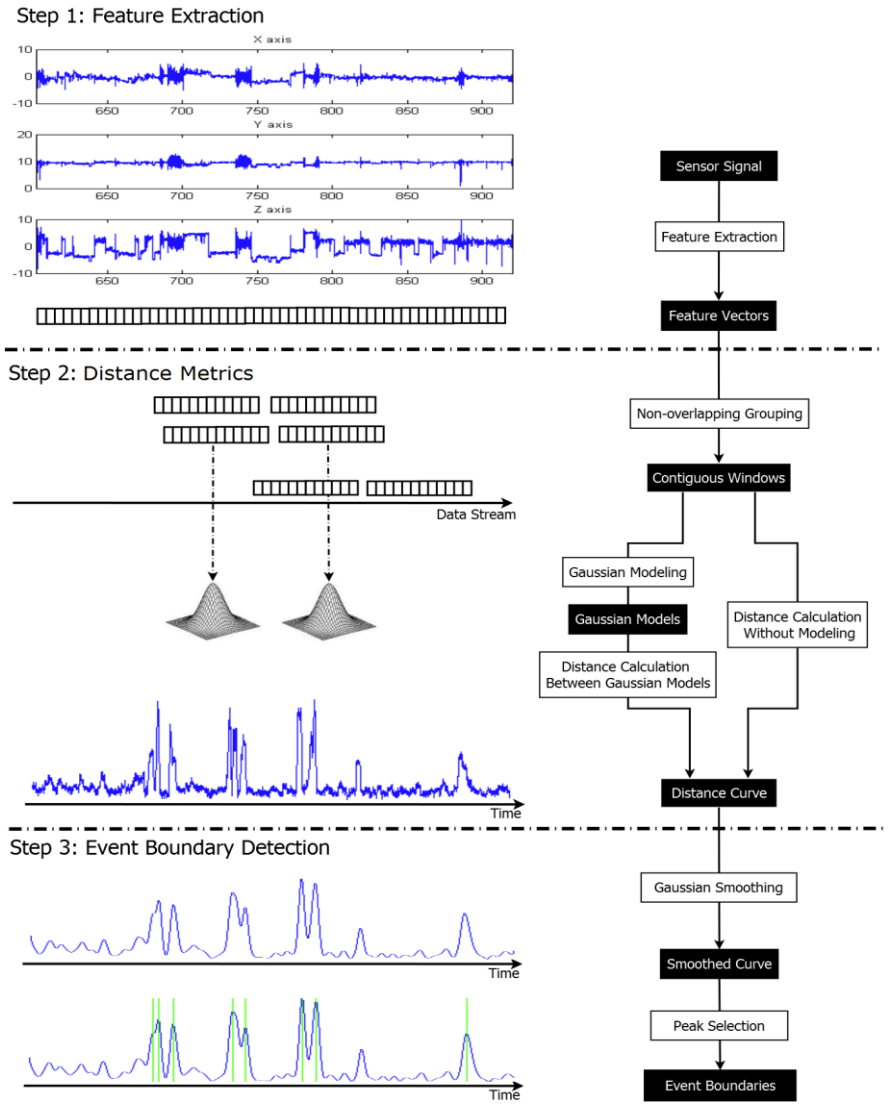
Motivation and Goals

- **사건 분절(event segmentation)**
  - 지속적으로 행해지는 사람의 일상 활동을 유의미한 사건들의 집합으로 나누는 작업

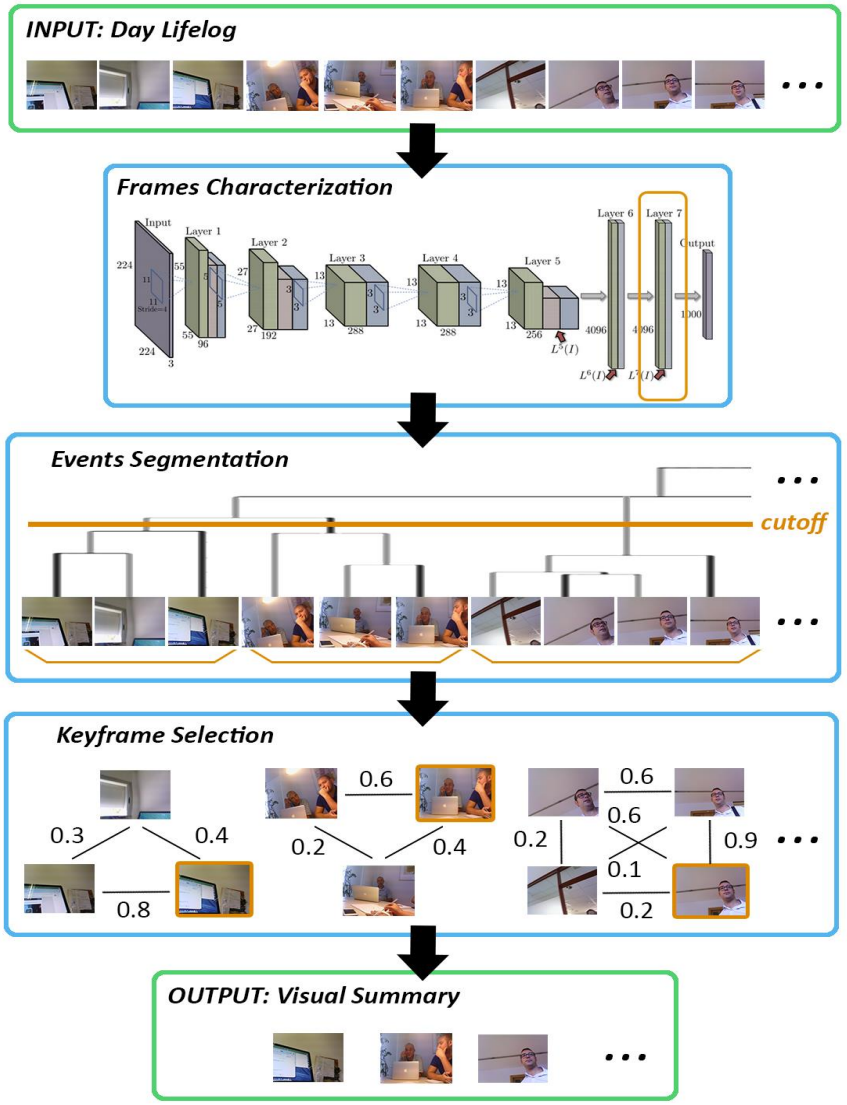
- **기존 방법의 한계**
  - 사건이 가지는 의미적인 상황 요소가 아닌 센서 데이터의 물리적인 변화가 발생하는 시점을 탐색
  - 특정 장비에 의존적이고, 센서 데이터에 내포된 노이즈에 취약하며, 보편적으로 사용되기 어려움



Related Works



Zhuang et al. (2013)



Bolanos et al. (2015)

Overview of Methodology

Life-logging data from wearable sensors

Image stream

Generating image descriptions by using CNN & B-RNN

Scene descriptions

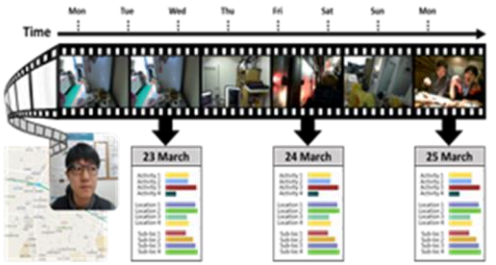
Sentence encoding by using Skip-thoughts model

Sentence vectors

Similarity measurement

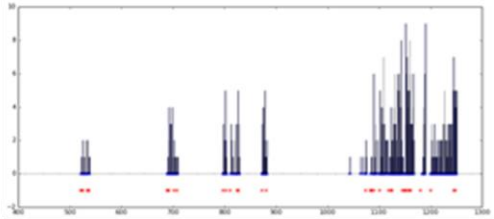
Unique event index

Event segmentation offsets



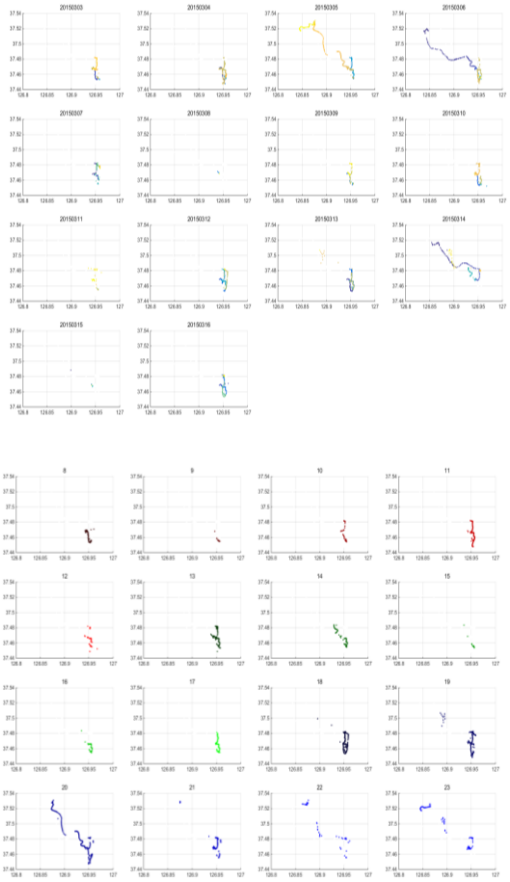
a person walking down a street with luggage
[-0.01716853 0.00315999 ..., 0.02567324]

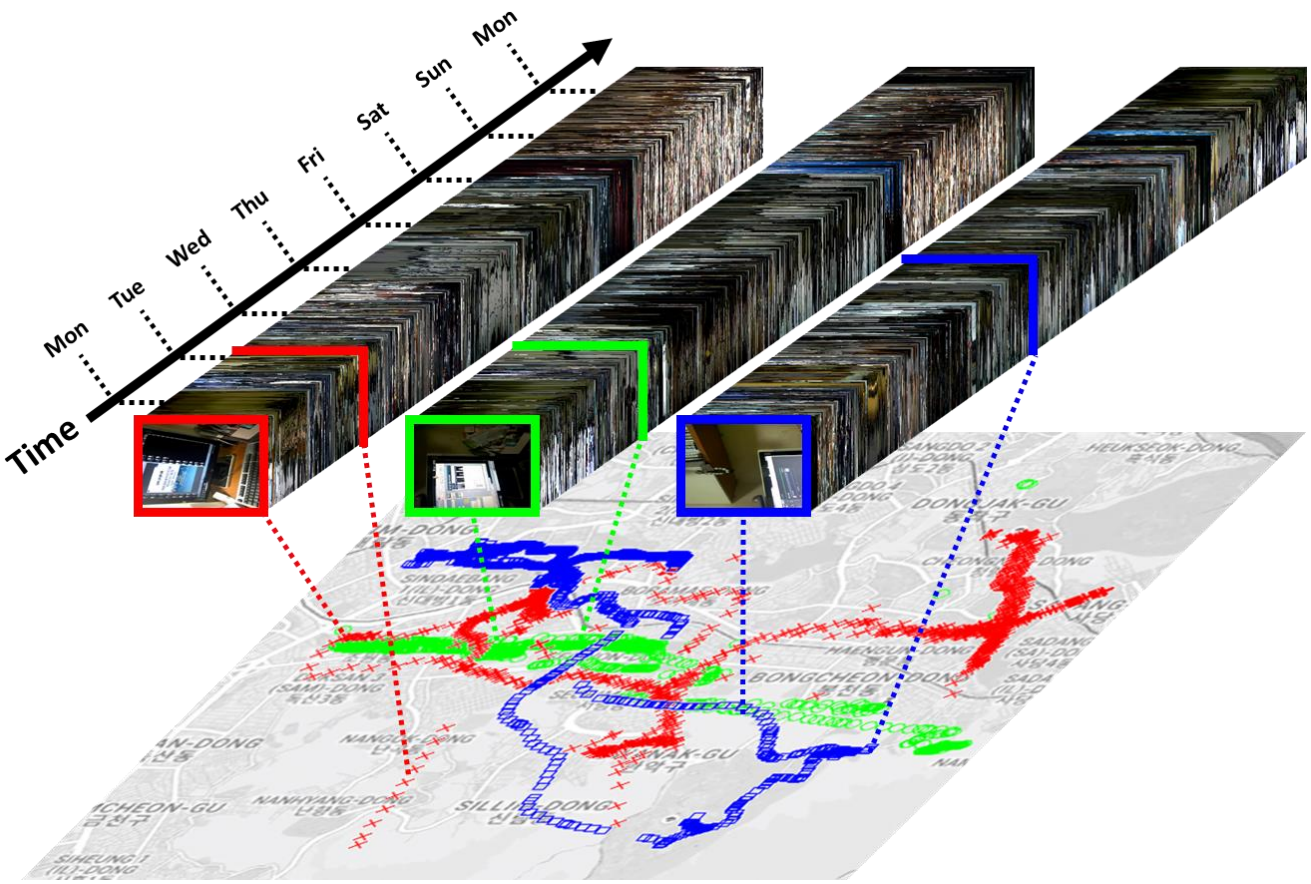
a person walking down a street with a suitcase
[-0.01748121 0.00682487 ..., 0.01648329]



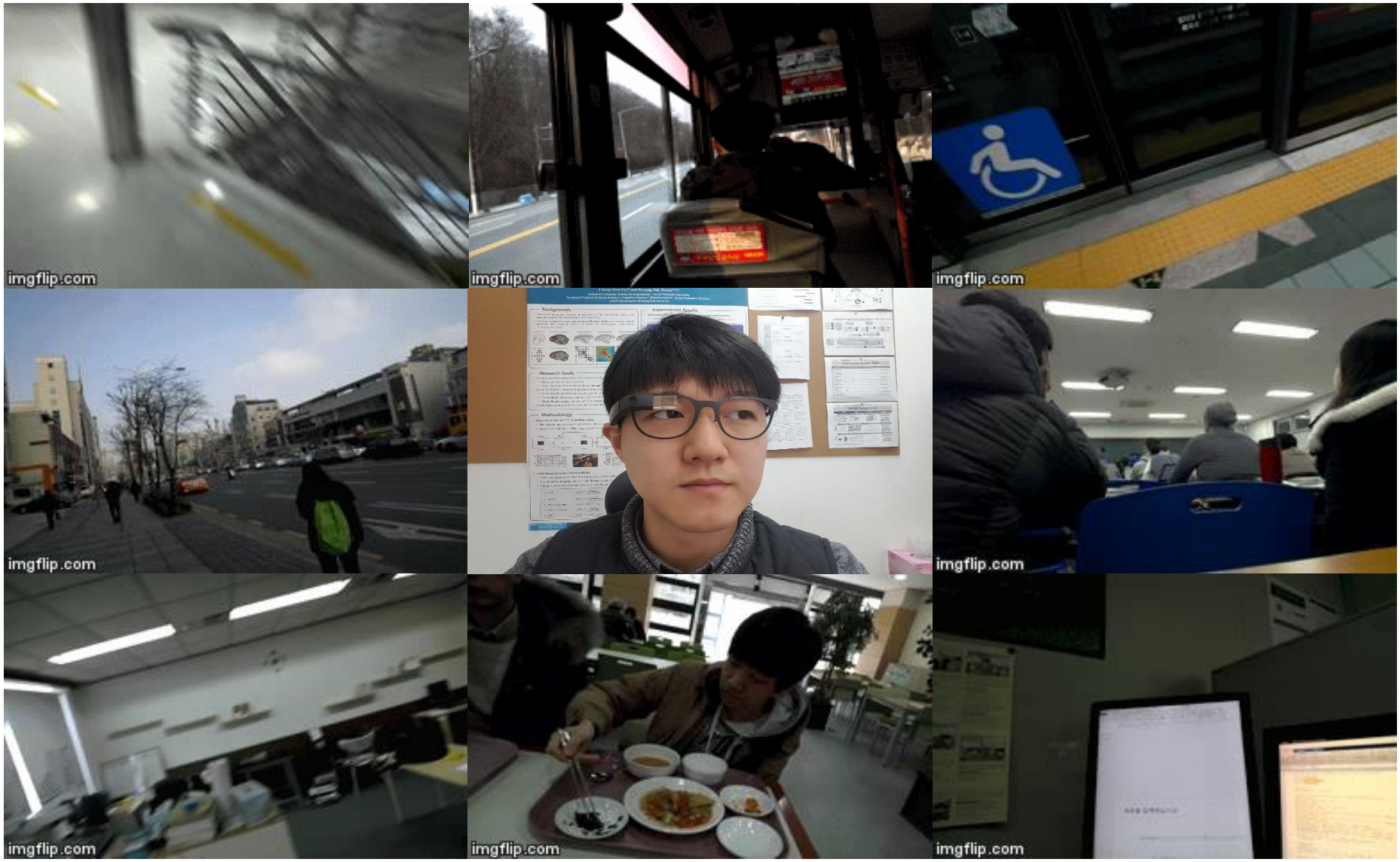
Egocentric Video Stream from Wild

- **Real-world lifelogs collected through the Google Glass for 46 days for 3 subjects**
- **Multi-modal sensory stream including image, audio, gyroscope, gps, etc.**



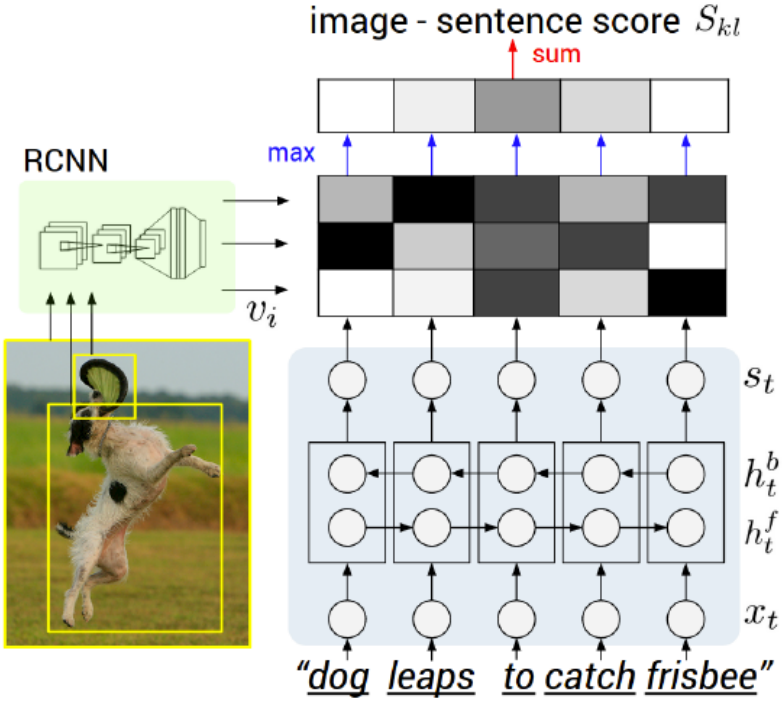


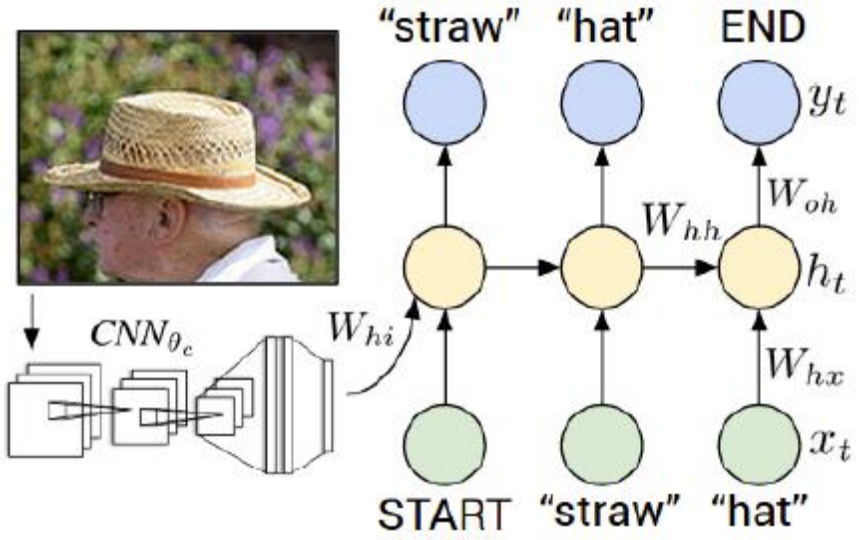
Egocentric Video Stream from Wild



Description Generation

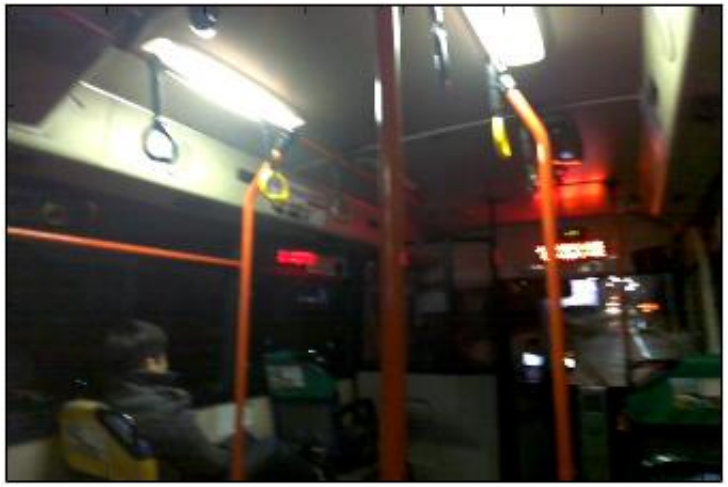
- **영상 서술 문장 생성 방법**
  - 라이프로깅 영상 스트림에 나타난 객체들을 RCNN (Region ConvNet, pretrained w/ ImageNet) 을 통해 인식
  - 객체들과 문장 사이의 관계를 학습한 BRNN (Bidirectional RNN)* 을 이용하여 영상을 설명하는 문장 생성





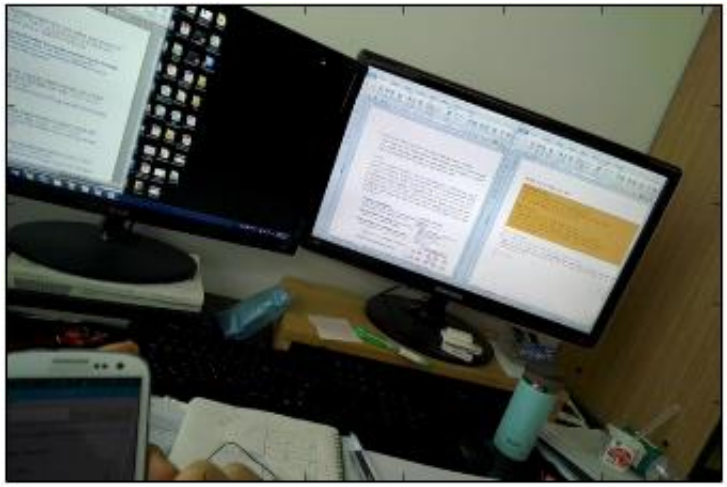
*Karpathy & Fei-Fei (2015), Deep Visual-Semantic Alignments for Generating Image Descriptions
https://github.com/karpathy/neuraltalk

Generated Descriptions



a man is standing in front of a bus

a man and a woman sitting at a table eating food



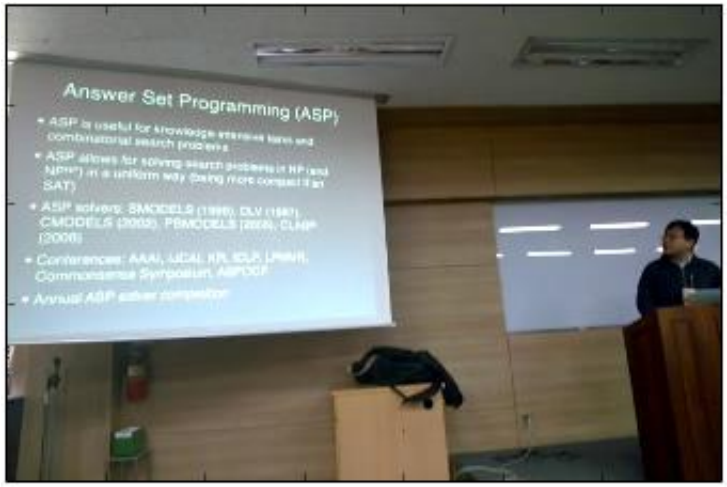
a desk with a computer monitor and keyboard

a car is driving down the street in a city

Generated Descriptions

a cell phone is sitting on a table

a book shelf filled with lots of books

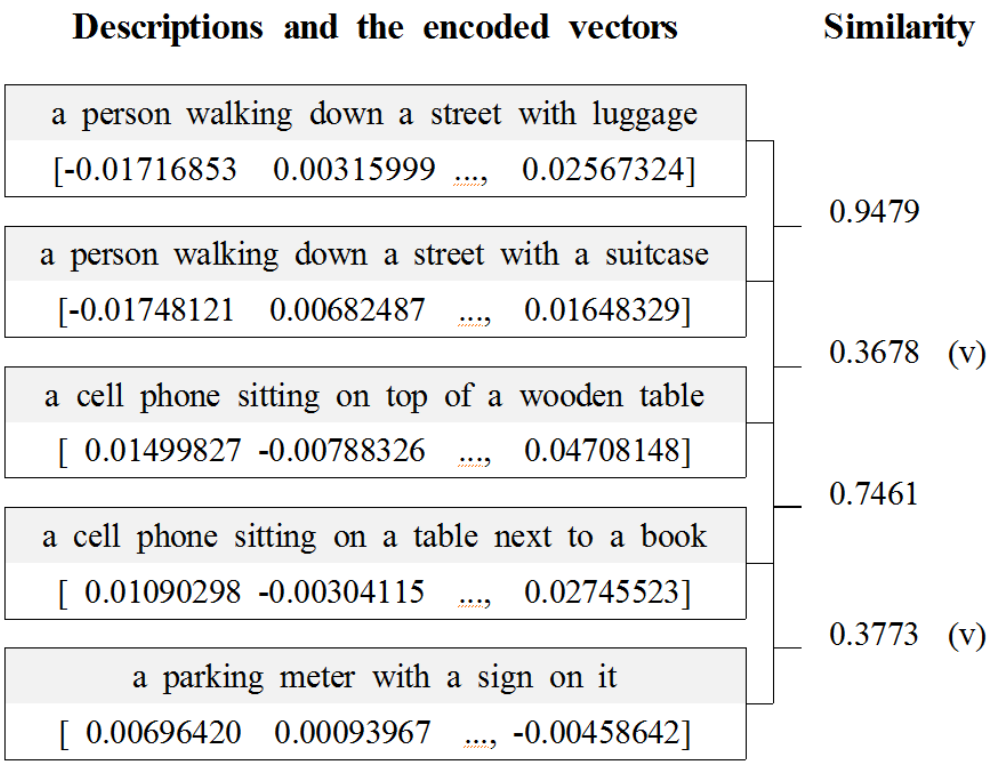


a man is standing in front of a tv

a parking meter on a city street with graffiti on it

Extracting Segmentation Offsets

- **문장 인코딩 및 유사도 계산을 통한 사건 분절 오프셋 생성**
  - 영상 설명문을 Skipthoughts model을 이용하여 벡터 형태로 인코딩
  - 인접 문장 벡터간 유사도를 Cosine similarity 계산을 통해 계산하여 사건 분절의 지표 도출



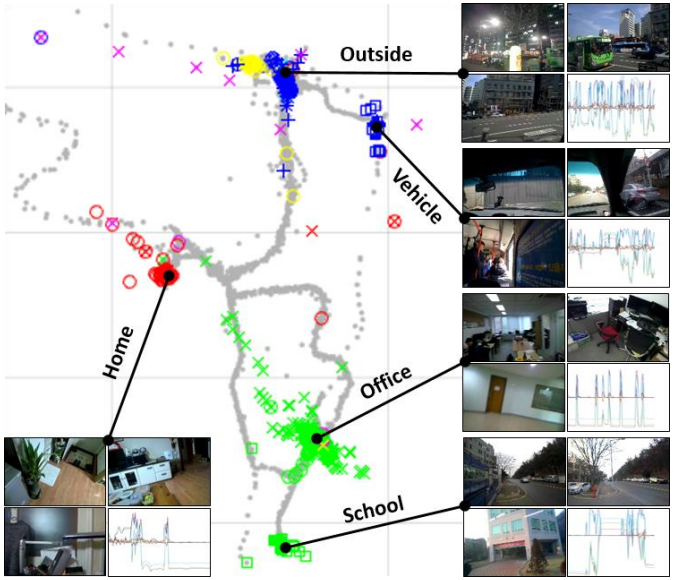
$$T = \frac{1}{2(N-1)} \sum_{i=1}^{N-1} (1 - Cos\,Sim(\overrightarrow{x_{i+1}}, \overrightarrow{x_i}))$$

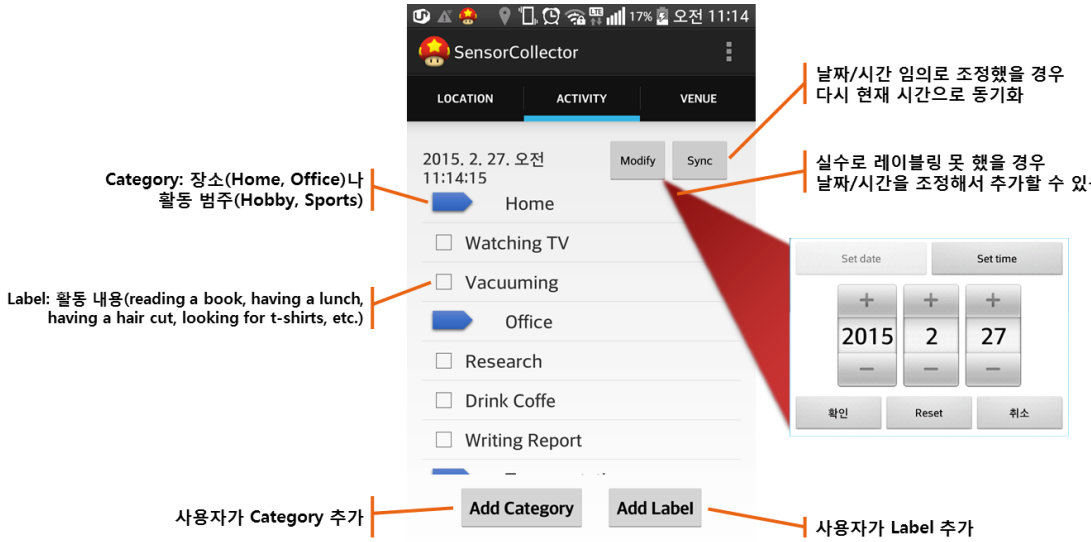
*Kiros et al. (2015), Skip-Thought Vectors
https://github.com/ryankiros/skip-thoughts

Experimental Results (1/4)

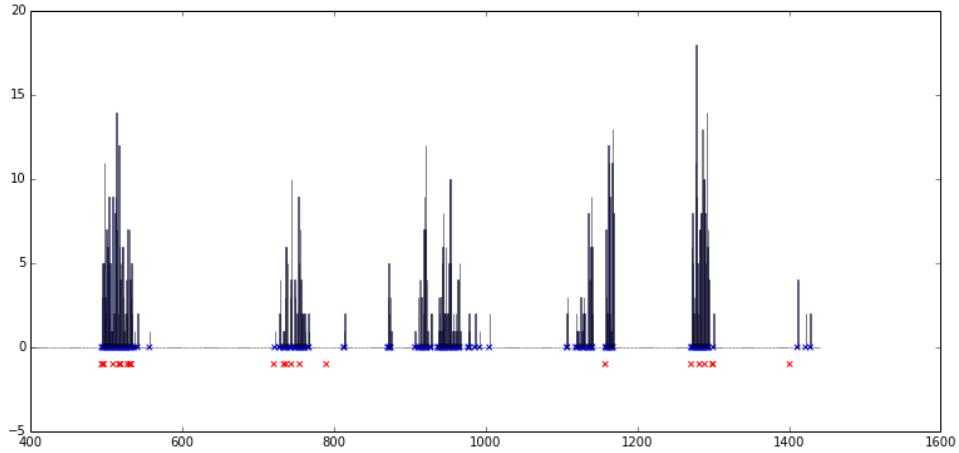
- **실험 데이터**
  - 약 150시간 분량의 라이프로깅 데이터 중 비디오 데이터와 레이블 사용
  - 2명의 실험 참가자가 각각 14일 동안 스마트 안경을 착용하고 생활하면서 수집

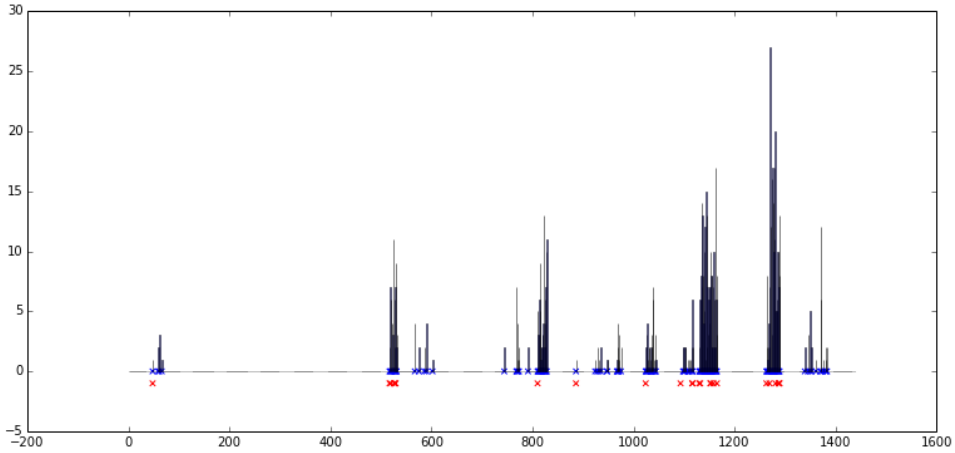
Subject	Video	Image	Activity Labels
Subject 1	95.11 hours	334,238	992
Subject 2	59.12 hours	205,819	497
Total	154.23 hours	540,057	1,489

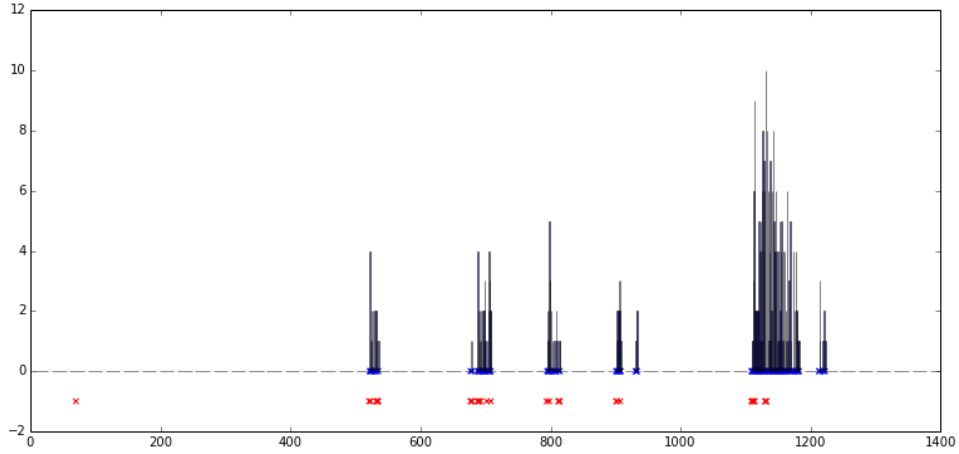


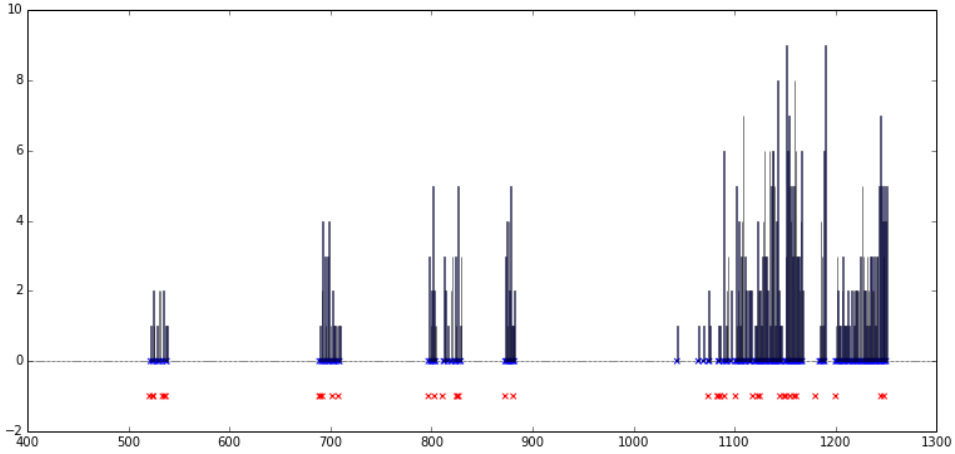


Experimental Results (2/4)

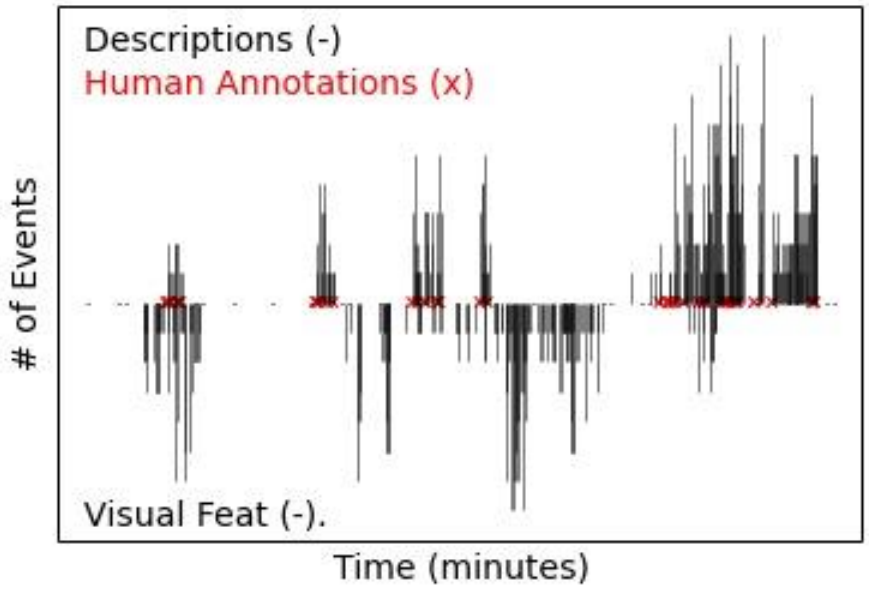








Experimental Results (3/4)



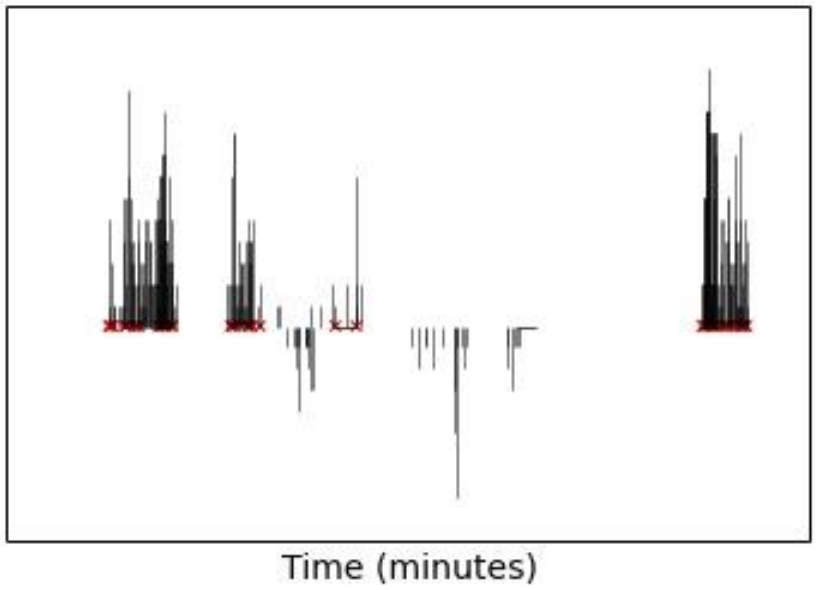
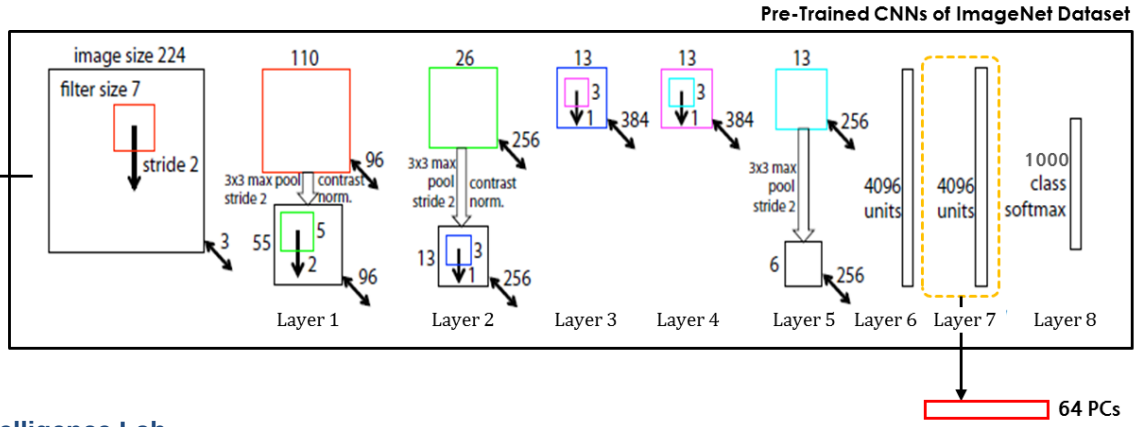
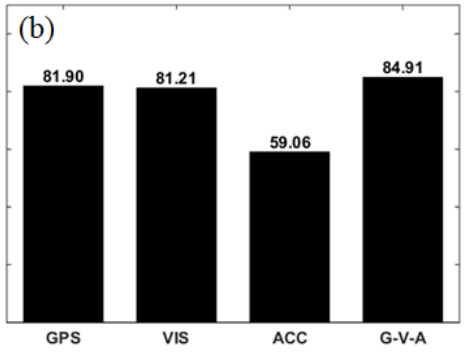
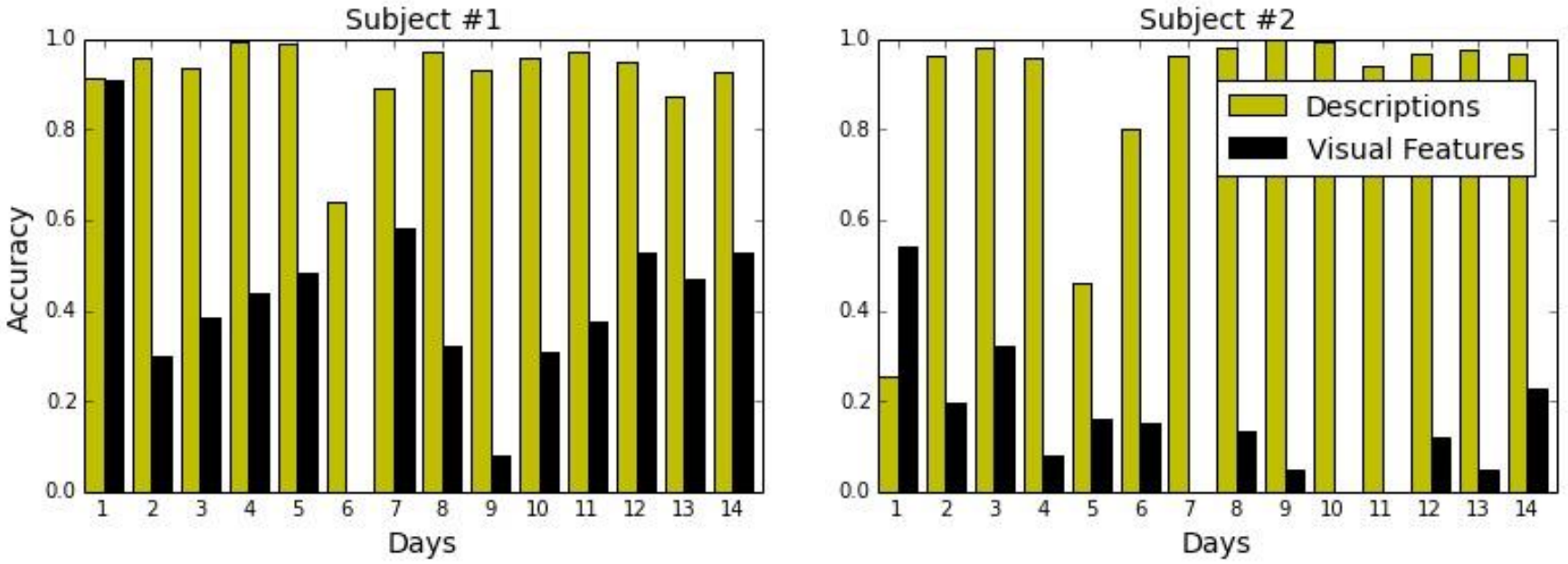



Image Stream





Experimental Results (4/4)



Test data	Description-based	Visual feature-based
Subject 1	92.158%	39.078%
Subject 2	87.119%	14.548%

Conclusions

- **Contributions**
  - 라이프로깅 영상의 자동 서술 생성을 통해 상황 정보가 반영된 사건 분절을 수행하는 새로운 방법을 제안
  - 실제 수집 데이터를 통해 제안한 방법을 검증 및 비교한 결과를 제시

- **Discussions**
  - 공간 단위 분절(RCNN)에 기반한 영상 서술을 통해 상황 정보가 반영된 것으로 해석
  - 효과를 강화하기 위해 여러 장의 영상을 함께 사용한 서술 문장 생성, 또는 다른 종류의 센서 데이터를 추가로 이용하여 보다 고차원의 상황 정보를 반영한 사건 분절을 시도 가능
  - 자동으로 분절된 사건들에 내포된 상황 정보에는 사람이 생각하지 못한 다양한 정보들이 더 남아있을 가능성이 있을 것으로 추측
  - 사건 분절 이후 서술문을 이용하여 해당 이벤트를 설명 가능할 것으로 판단됨
  - 계산량이 너무 많음 → Attention mechanism, 영상 처리 기법 등 이용

THANK YOU