

라이프로깅 영상 스트림의 서술 생성을 이용한 자동 사건 분절

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

이충연, 한동식, 장병탁

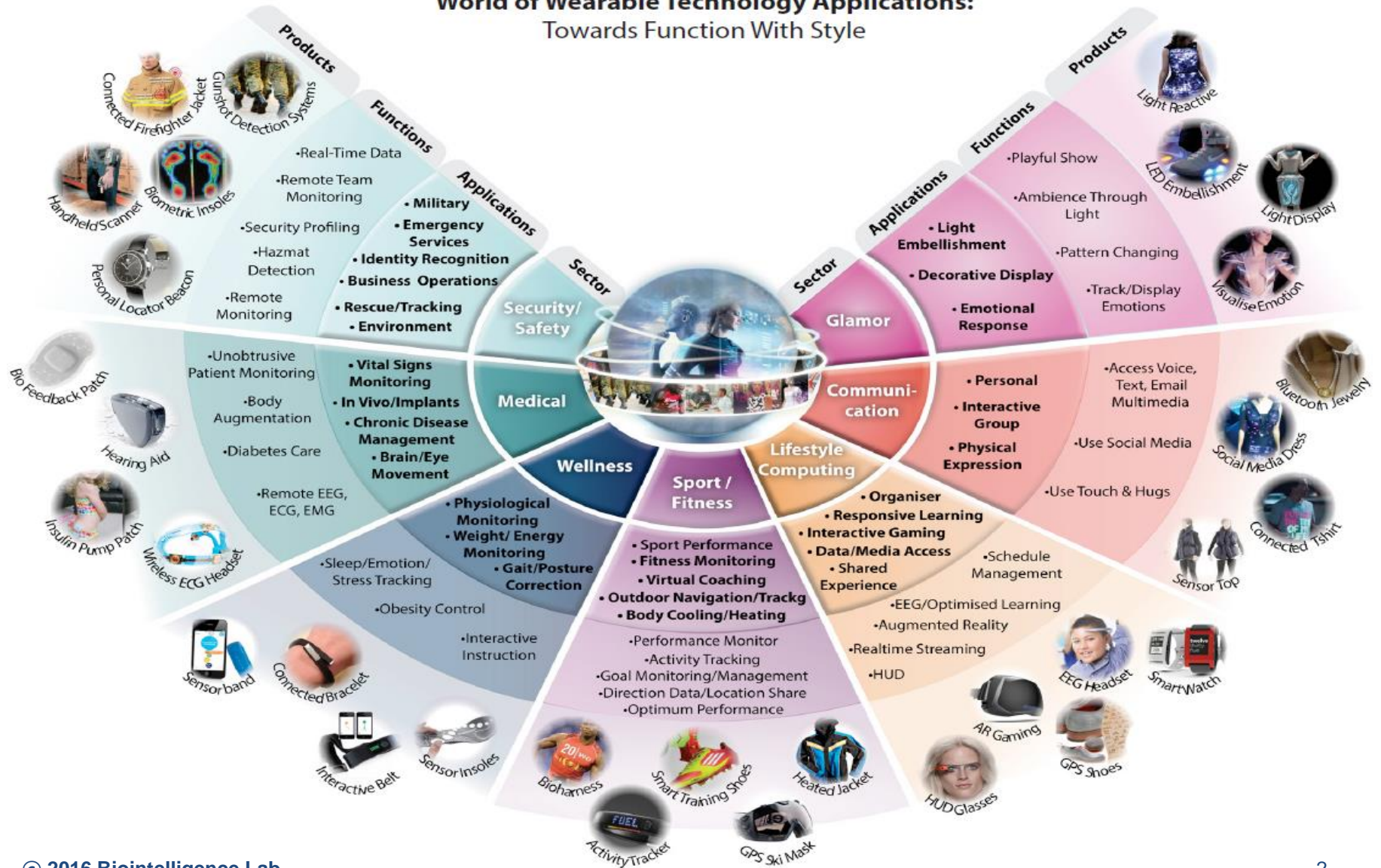
한국컴퓨터종합학술대회(KCC)

2016. 7. 1

서울대학교 컴퓨터공학부 바이오지능연구실

Motivation and Goals

World of Wearable Technology Applications: Towards Function With Style



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, 불편적으로 사용되기 어려움

A day's SenseCam images (3,000 – 4,000)



Event Segmentation

Multiple Events



Finishing work
in the lab

At the bus
stop

Chatting at Skylon Hotel
lobby

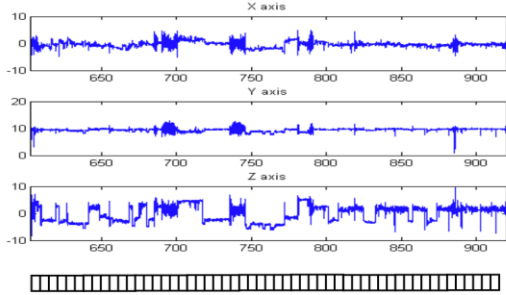
Moving to a
room

Tea time

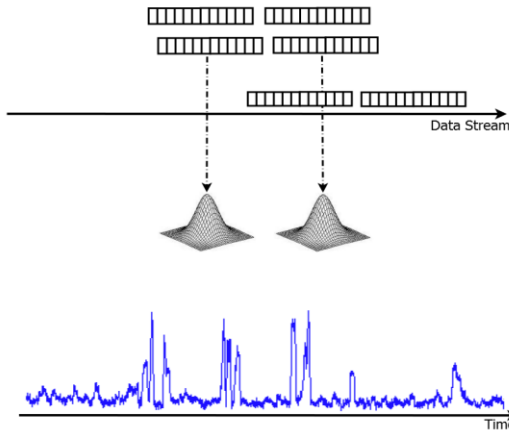
On the way
back home

Related Works

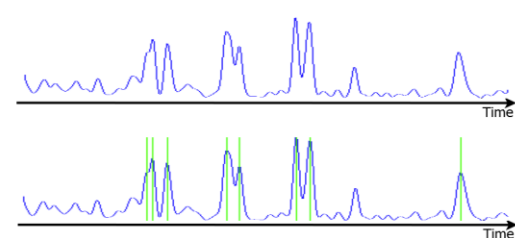
Step 1: Feature Extraction



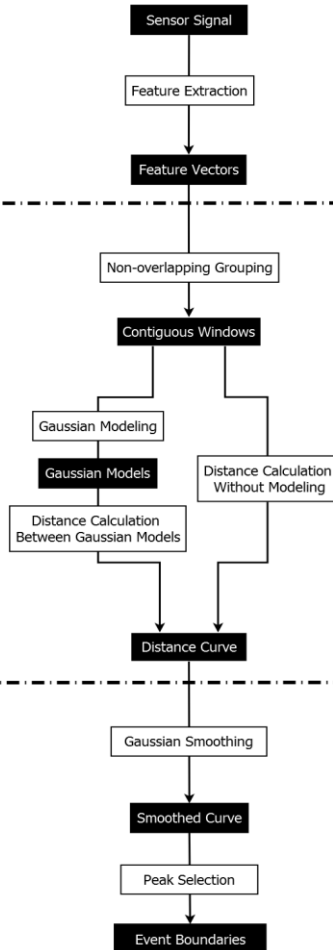
Step 2: Distance Metrics



Step 3: Event Boundary Detection



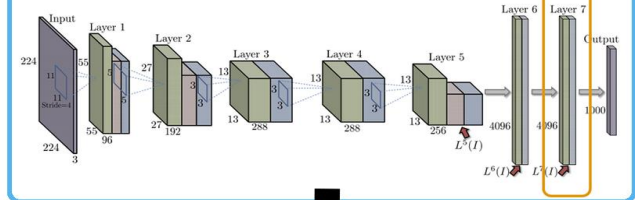
Zhuang et al. (2013)



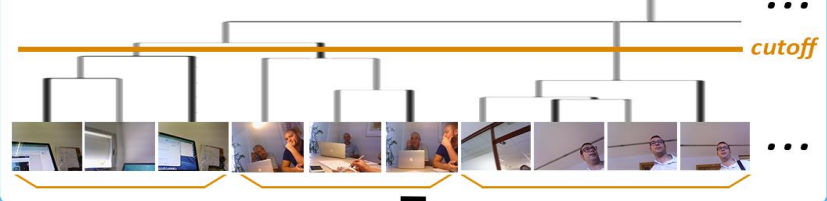
INPUT: Day Lifelog



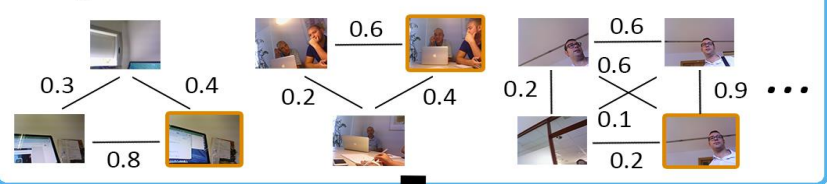
Frames Characterization



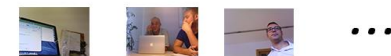
Events Segmentation



Keyframe Selection

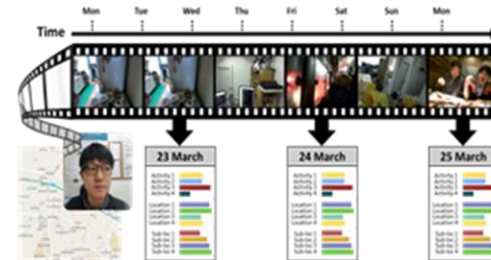
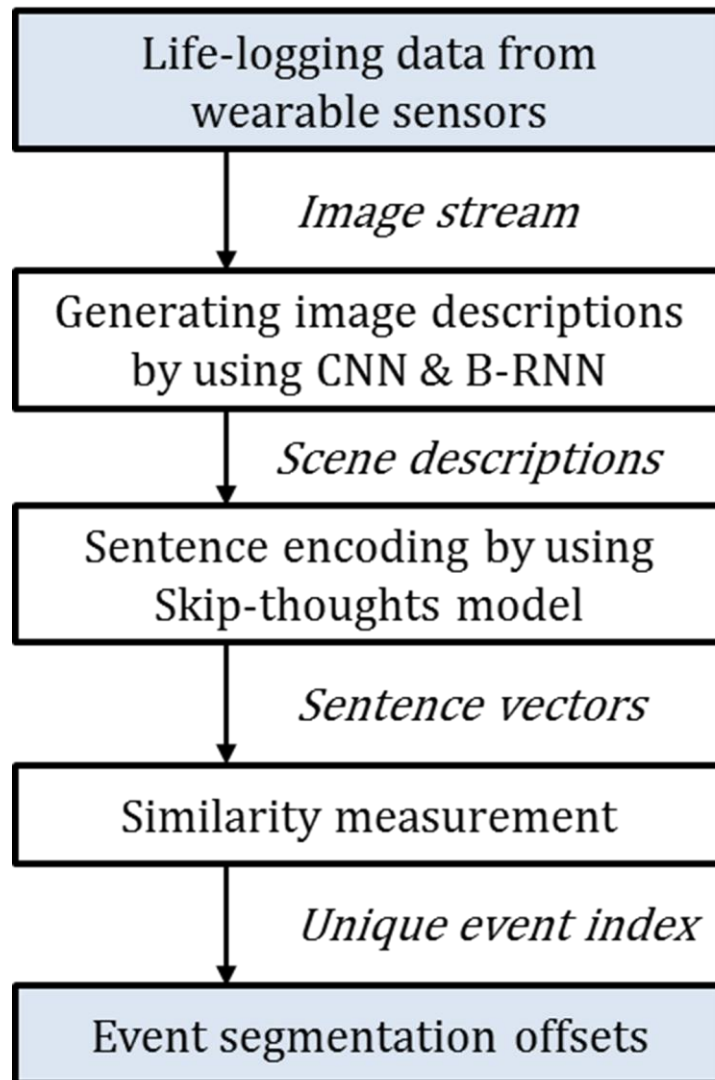


OUTPUT: Visual Summary



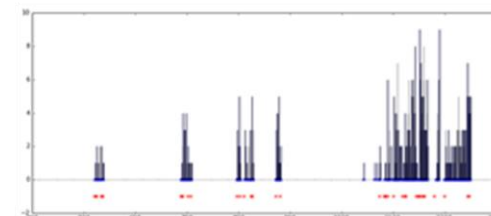
Bolanos et al. (2015)

Overview of Methodology



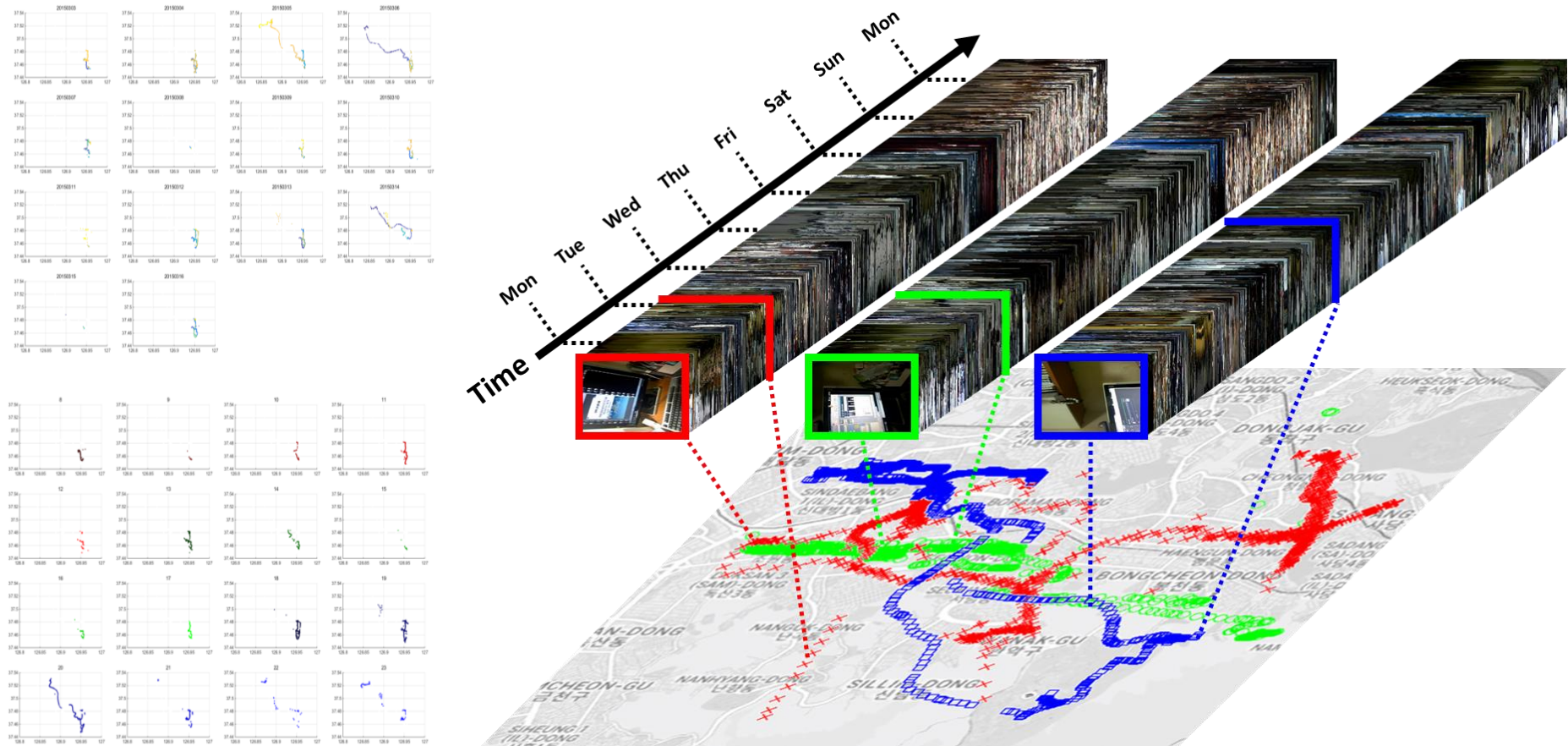
a person walking down a street with luggage
[-0.01716853 0.00315999 ..., 0.02567324]

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

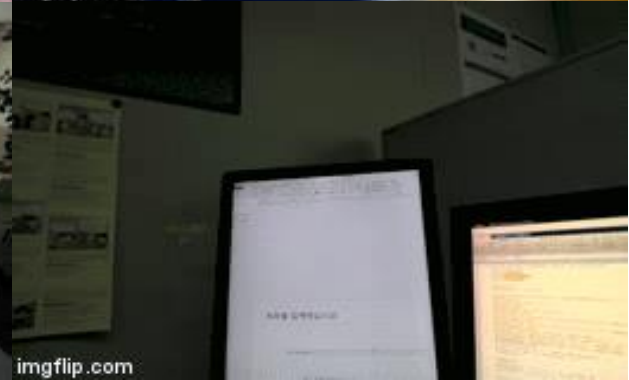
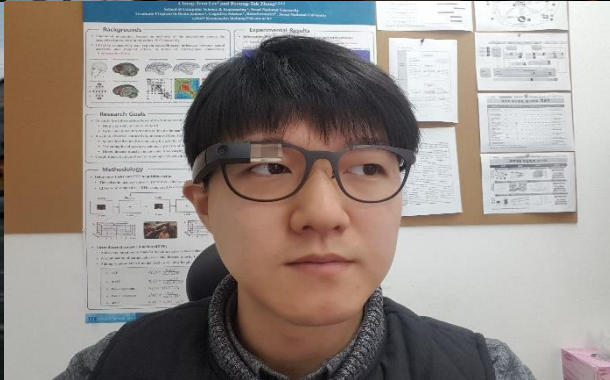


Egocentric Video Stream from Wild

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



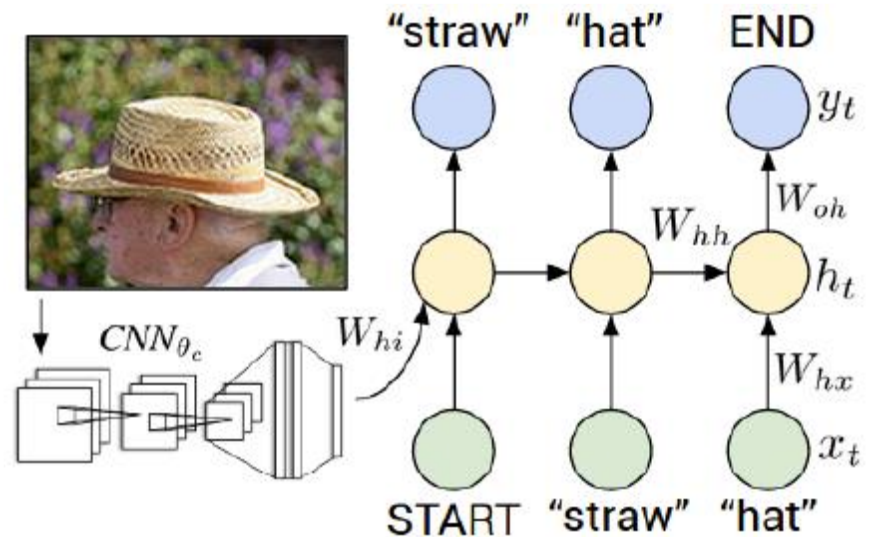
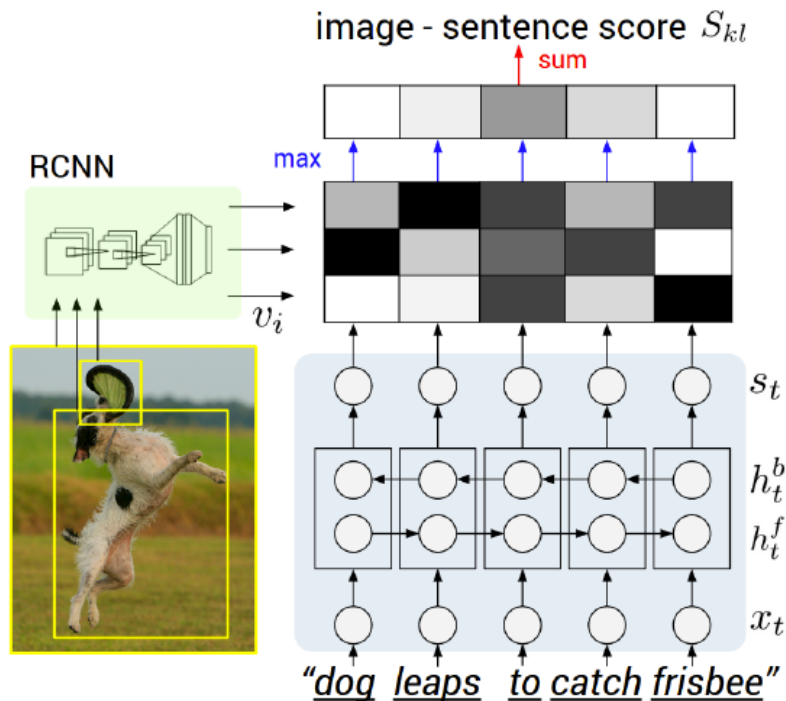
Egocentric Video Stream from Wild



Description Generation

■ 영상 서술 문장 생성 방법

- 라이프로그 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>

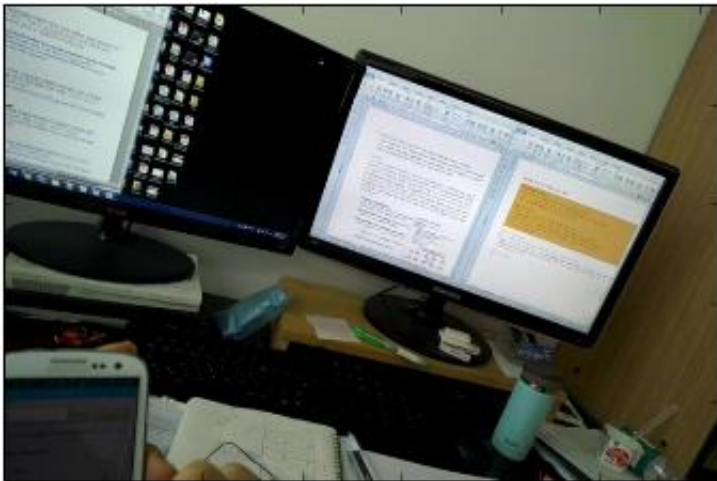
Generated Descriptions



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

Descriptions and the encoded vectors	Similarity
<div>a person walking down a street with luggage</div> <div>[-0.01716853 0.00315999 ..., 0.02567324]</div>	0.9479
<div>a person walking down a street with a suitcase</div> <div>[-0.01748121 0.00682487 ..., 0.01648329]</div>	
<div>a cell phone sitting on top of a wooden table</div> <div>[0.01499827 -0.00788326 ..., 0.04708148]</div>	0.3678 (v)
<div>a cell phone sitting on a table next to a book</div> <div>[0.01090298 -0.00304115 ..., 0.02745523]</div>	0.7461
<div>a parking meter with a sign on it</div> <div>[0.00696420 0.00093967 ..., -0.00458642]</div>	0.3773 (v)

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

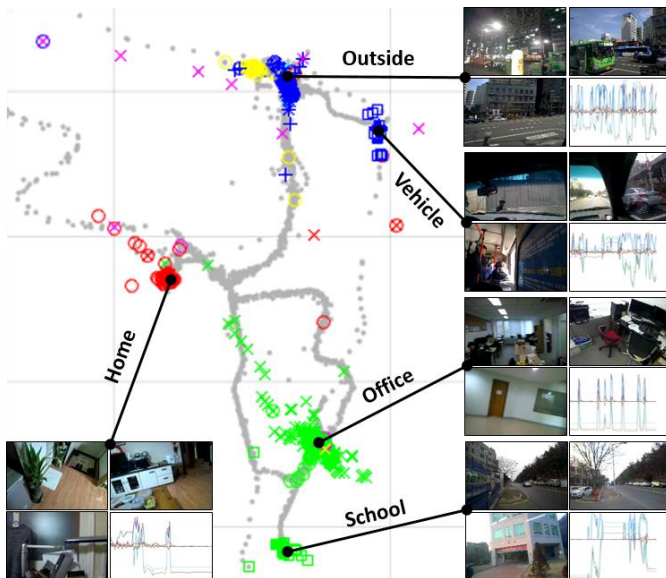
*Kiros et al. (2015), Skip-Thought Vectors
<https://github.com/ryankiros/skip-thoughts>

Experimental Results (1/4)

■ 실험 데이터

- 약 150시간 분량의 라이프로그 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



Category: 장소(Home, Office)나 활동 범주(Hobby, Sports)

Label: 활동 내용(reading a book, having a lunch, having a hair cut, looking for t-shirts, etc.)

날짜/시간 임의로 조정했을 경우 다시 현재 시간으로 동기화

실수로 레이블링 못 했을 경우 날짜/시간을 조정해서 추가할 수 있음

사용자가 Category 추가

사용자가 Label 추가

SensorCollector

LOCATION ACTIVITY VENUE

2015. 2. 27. 오전 11:14:15

Modify Sync

Home

☐ Watching TV

☐ Vacuuming

☒ Office

☐ Research

☐ Drink Coffee

☐ Writing Report

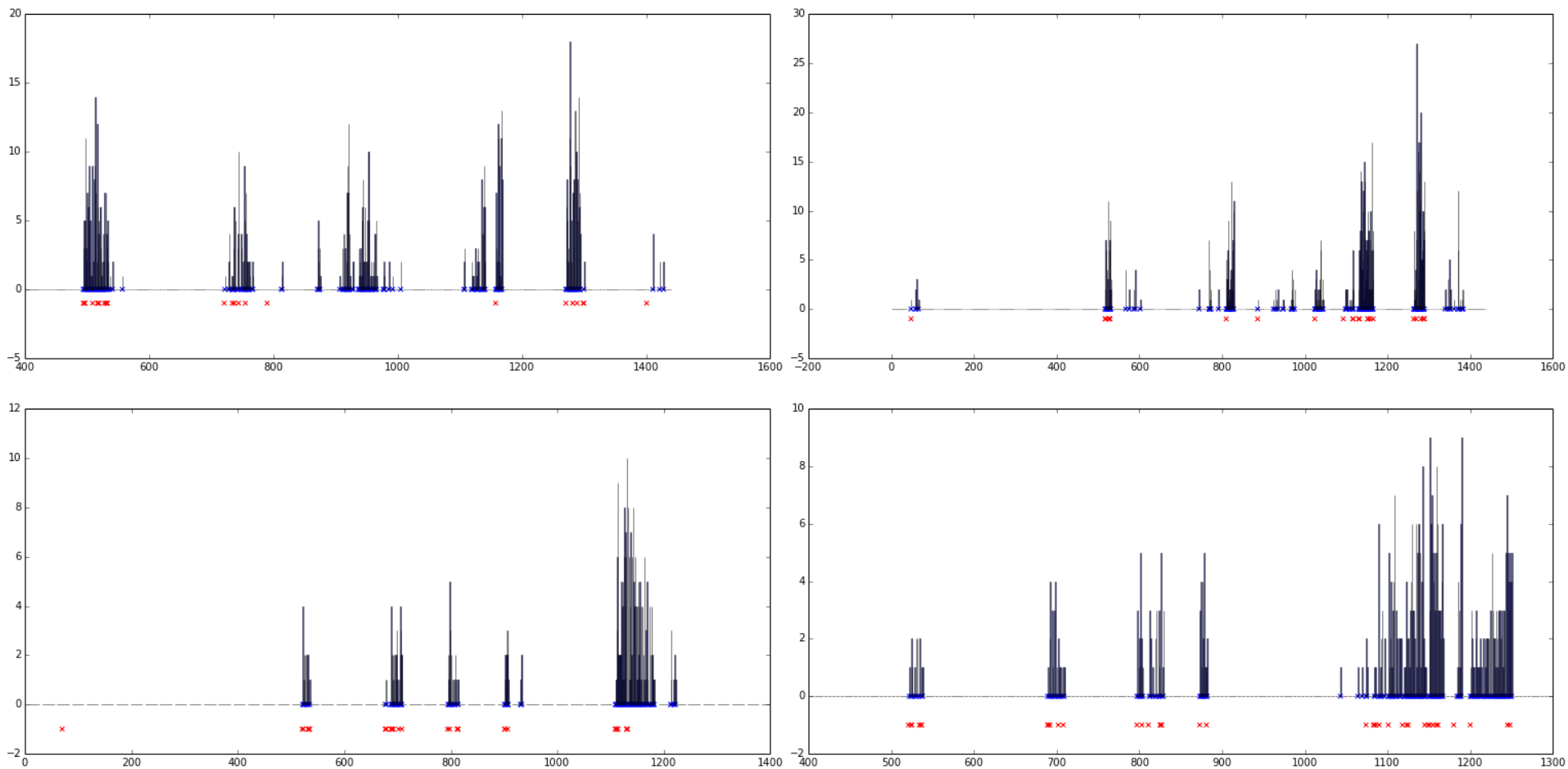
Add Category Add Label

Set date Set time

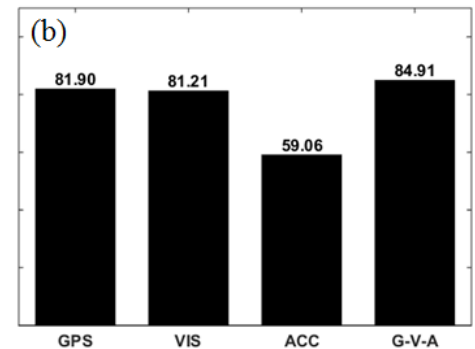
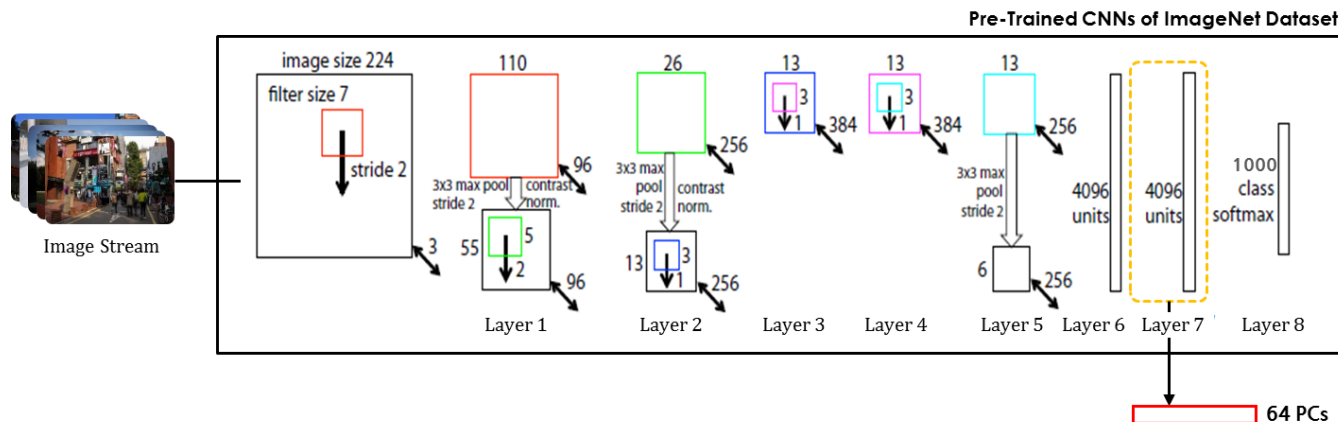
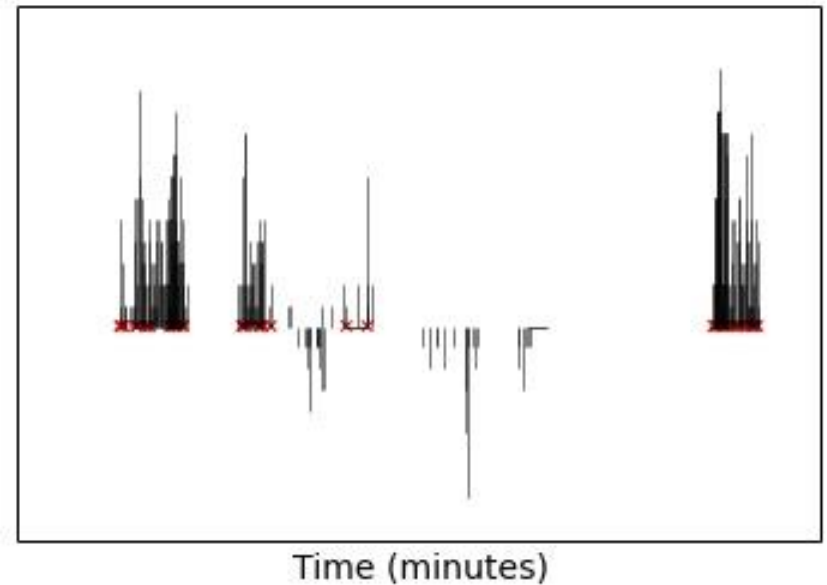
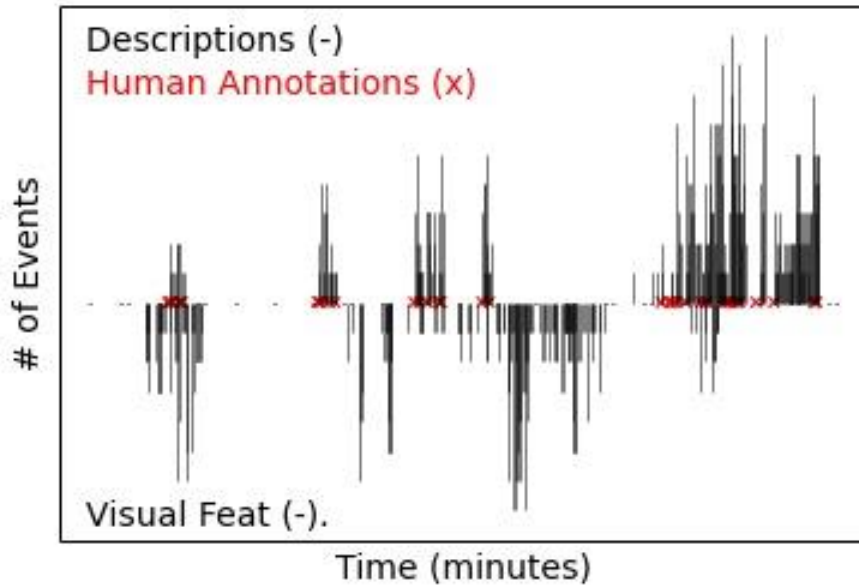
2015 2 27

확인 Reset 취소

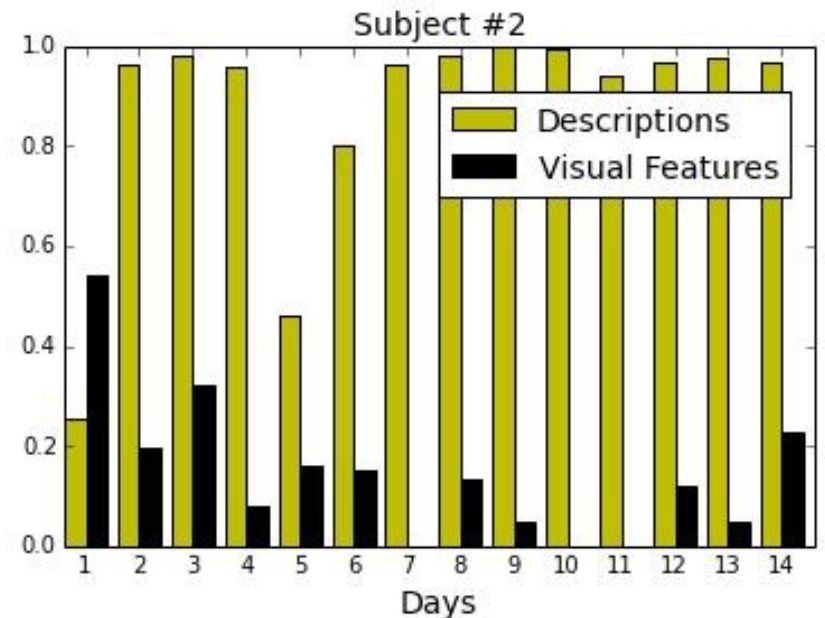
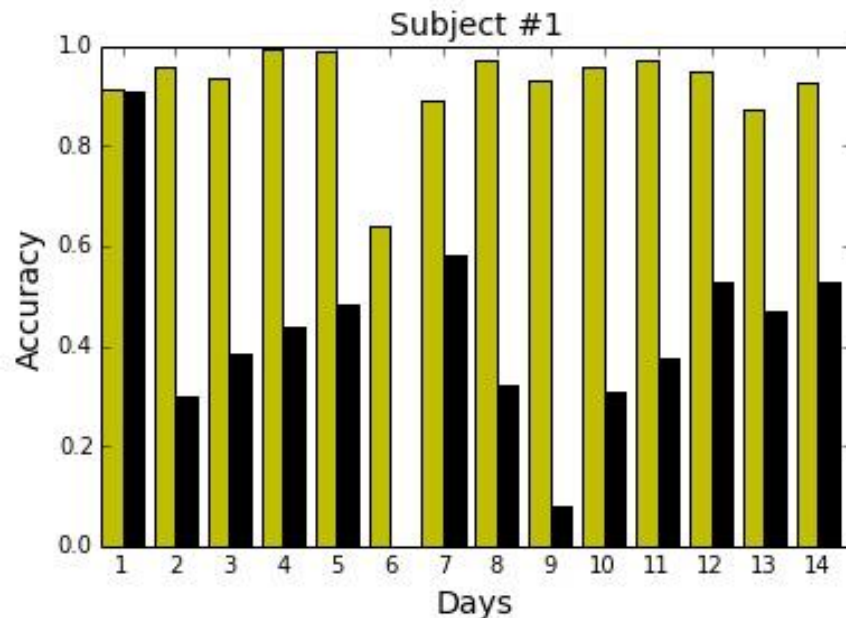
Experimental Results (2/4)



Experimental Results (3/4)



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

- 라이프로그 영상 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