Lasagna: Towards Deep Hierarchical Understanding and Searching over Mobile Sensing Data

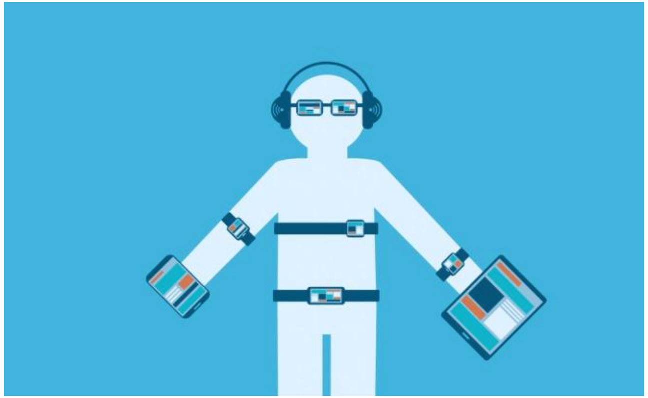
Cihang Liu, Lan Zhang, Zongqian Liu, Kebin Liu, Xiangyang Li, Yunhao Liu
Tsinghua University,University of Science and Technology of China

Outline

1. Background
2. State-of-the-Art
3. Deep and Hierarchal Understanding of Mobile Sensing Data
4. Semantic Based Activity Search
5. Implementation & Evaluation
6. Conclusion & Open Issues

1. Background

The Fascinating Smart Wearables



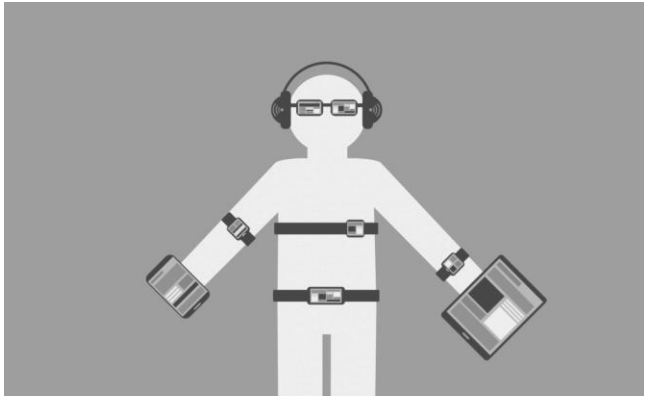
Market of smart wearables:

- 2016: $30bn
- 2018: $40bn
- 2023: $100bn

Promising Industries:

- Healthcare & Medical
- Fitness & Wellness
- Commercial
- Military…

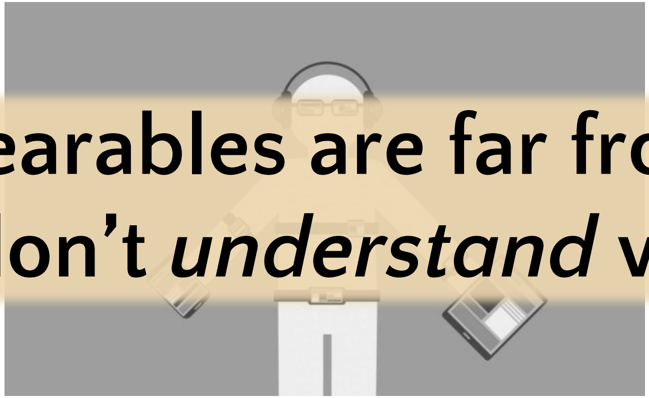
The Unsatisfying Smart Wearables



What wearables can do:

- Step counting
- Step counting
- …

The Unsatisfying Smart Wearables

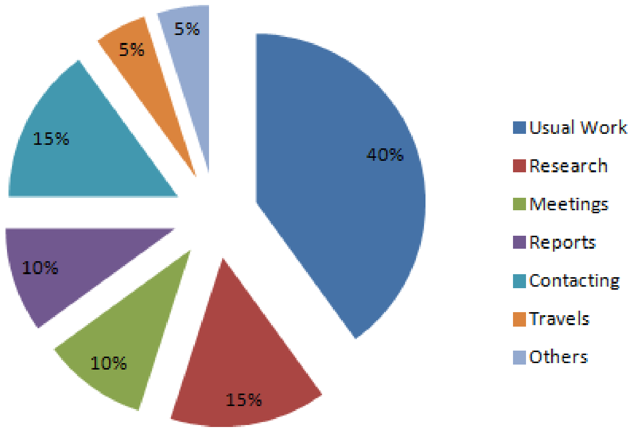


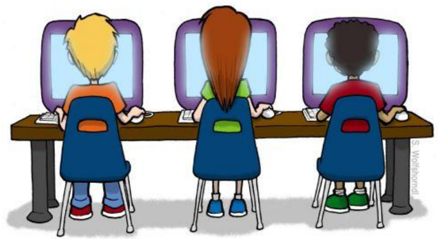
What wearables can do:

- Step counting
- Step counting
- …

Wearables are far from *smart* because they don't *understand* what we do everyday

Potential Applications





- Keep a smart diary of our daily activities
- Achieve accurate working performance calculation

Potential Applications

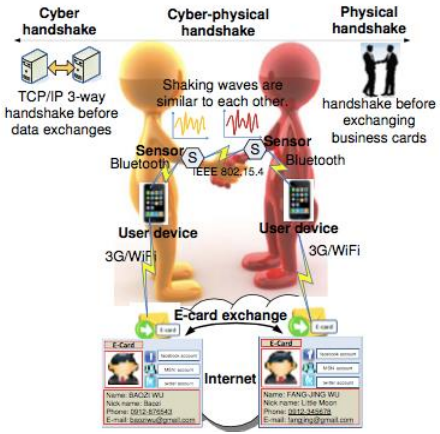
- Investigate civil health condition
- Study the cause of common occupational diseases

Lasagna Makes Wearables Smart

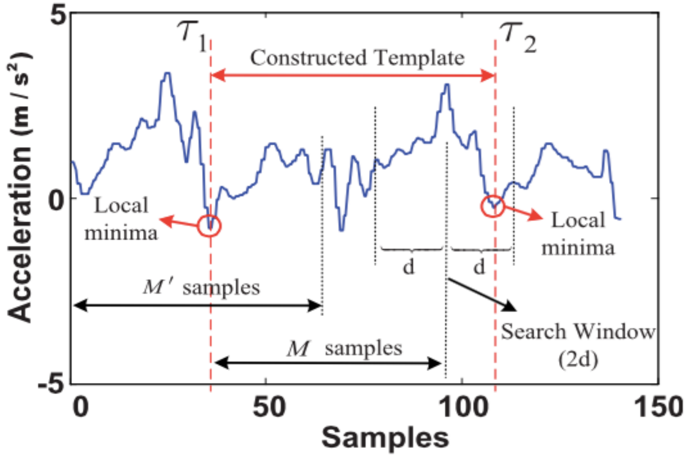
- Proposes deep hierarchical understanding of mobile sensing data
- Enables Semantic Based Activity Search (SBAS)

2. State-of-the-Art

Physical Model based Methods

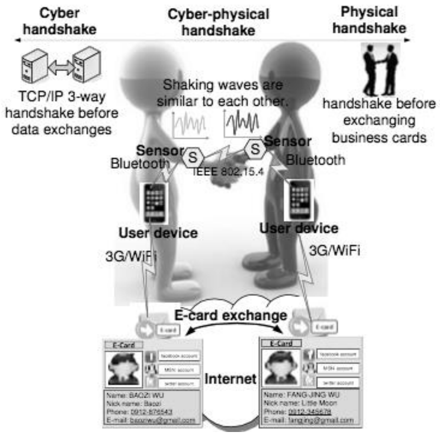


Handshake Model
(SIGCOMM ’11, poster)

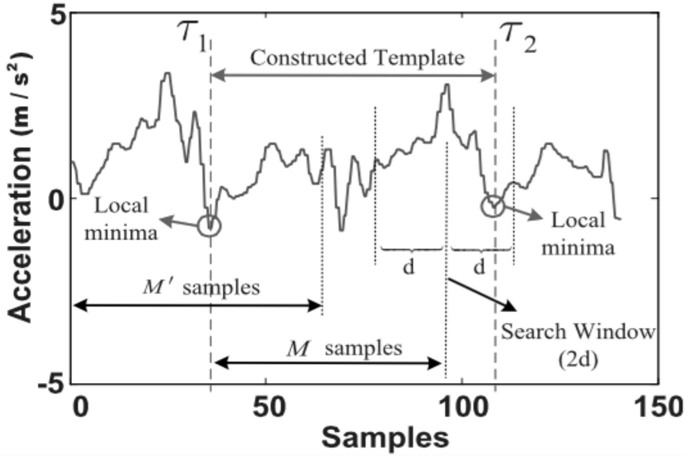


Gait Model
(TMC ‘13)

Physical Model based Methods



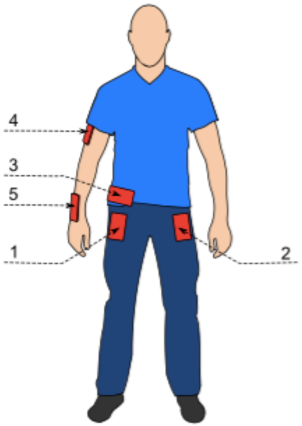
Handshake Model
(SIGCOMM '11, poster)



Gait Pattern
(TMC '13)

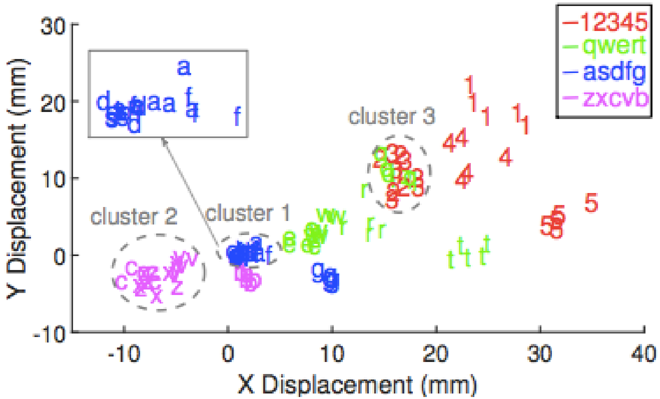
Targeting specific activities
Hard to spread to others

Feature Set based Methods



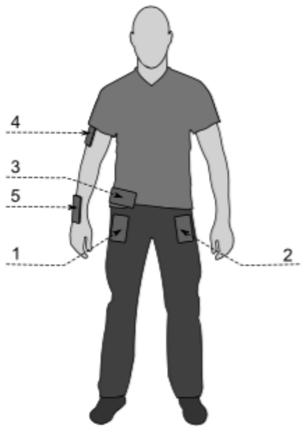
Feature Set	Associated Features
FS1	Mean
FS1	Standard Deviation
FS2	Median
FS2	Variance
FS2, FS3	Zero crossings
FS2, FS3	Root mean square value
FS4	Sum off FFT coefficients
FS4	Signal energy

Various motion sensors with different feature sets (Sensors ‘14)

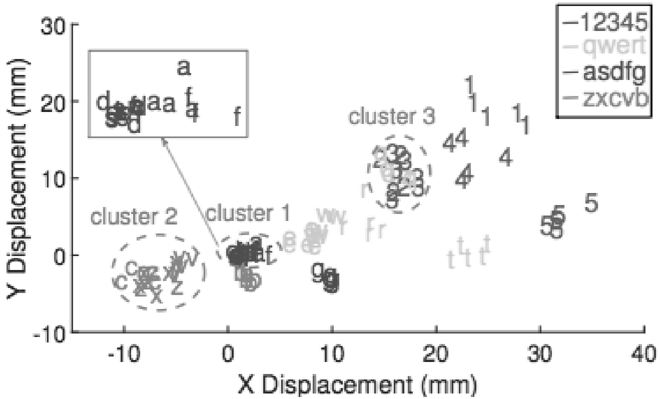


Mole (Mobicom ‘15)

Feature Set based Methods



Feature Set	Associated Features
FS1	Mean
FS1	Standard Deviation
FS2	Median
FS2	Variance
FS2, FS3	Zero crossings
FS2, FS3	Root mean square value
FS4	Sum off FFT coefficients
FS4	Signal energy

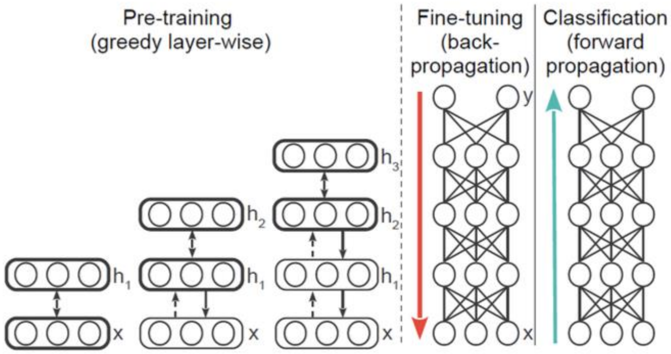


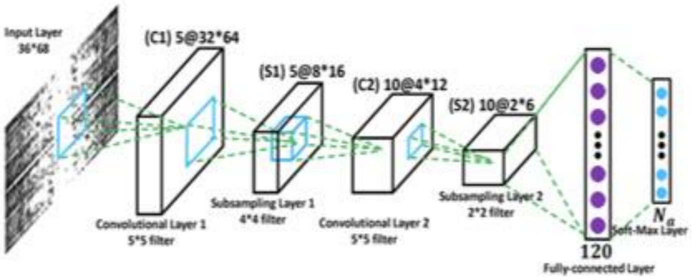
Various motion sensors with different feature sets (Sensors '14)

Mole (Mobicom '15)

Adopt statistical features
Cannot provide satisfying results

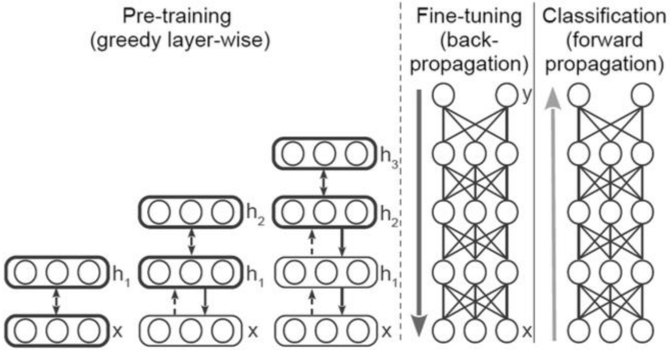
Supervised Deep Learning based Methods

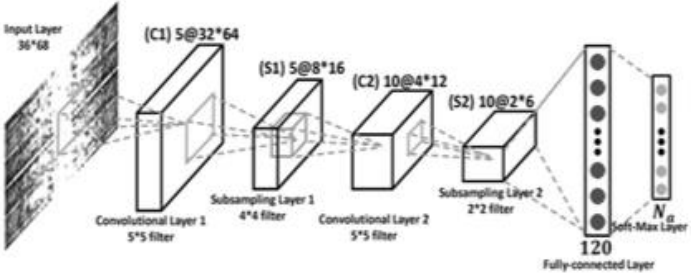




- DNN benefits the accuracy and robustness. (HotMobile '15)
- Using CNN and SVMs, features provide around 98% recognition accuracy. (ACM MM '15)

Supervised Deep Learning based Methods





Requires too much training data, training time and computation resource.

DNN benefits the accuracy and robustness. (HotMobile '15)

Using CNN and SVMs, features provide around 98% recognition accuracy. (ACM MM '15)

Challenges

Activity

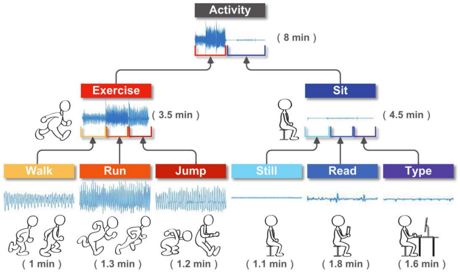
- Human activities are **arbitrary**, and rich in **hierarchical** semanteme.

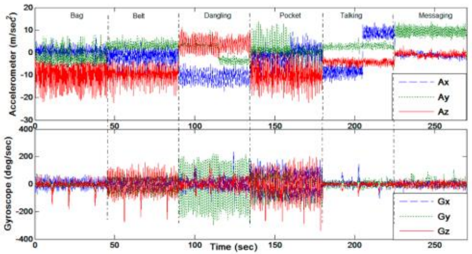
Data

- Data can be easily affected **diversities**.
(device, people, timescale, etc.)

Resource

- COTS devices are limited in resources.
(battery , computation, etc)





3. Deep Hierarchical Understanding of Mobile Sensing Data

Q1: How to describe arbitrary activities?

Basis

Inspiration: Basis

- A ***basis*** of a vector space **V** over a field ***F*** is a linearly independent subset of **V** that spans **V**.
- ***Spanning Poperty***:

For every $\mathbf{x}$ *in* **V**, *it is possible to choose* $a_1, \ldots, a_n \in \mathbf{F}$, *such that*

$$\mathbf{x} = a_1v_1 + \ldots + a_nv_n$$

Inspiration: Basis

- For two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$,

$$\boldsymbol{x}_1 = a_1 v_1 + \ldots + a_n v_n$$
$$\boldsymbol{x}_2 = a_1 v_1 + \ldots + a_n v_n$$

An arbitrary point can be represented by the basis.

Any two points are comparable according to the embedding (coordinates).

Convolution Kernel

Inspiration: Convolution Kernel

- Convolution kernels have been widely used in extracting the *latent* information.
- Different kernels can reveal different characteristics.

Edge $\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$

Sharpen $\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$

Gaussian Blur $\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$

Box Blur $\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

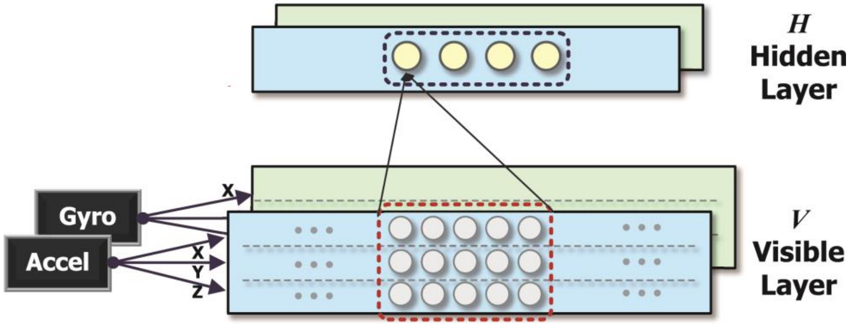
Idea: Adopt kernels as Motion Basis

- 1. Use ***diverse convolution kernels*** to reveal the characteristics of human activities.
- 2. ***Combine*** kernels as ***motion basis*** to get comprehensive understanding.

An arbitrary activity can be represented by the basis

Two activities are comparable according to the embedding.

CRBM Learns Motion Basis



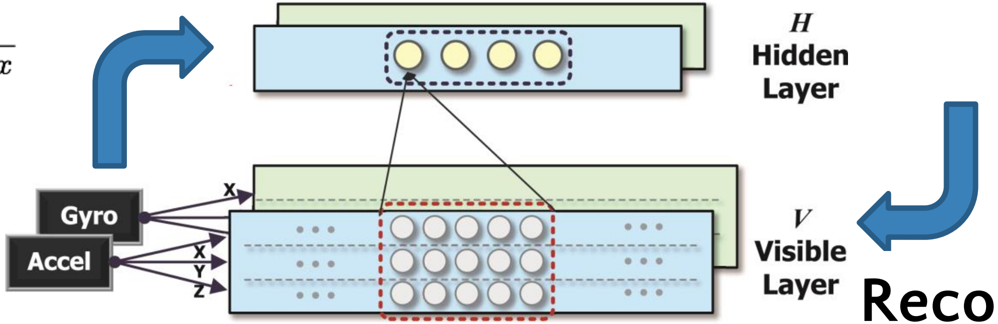
Convolution Restricted Boltzmann Machine

CRBM Learns Motion Basis

Inference

$$P(\mathbf{h}_{k,j} = 1|I) = \sigma((\widetilde{W}^k *_v I)_j + b_k),$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Reconstruction

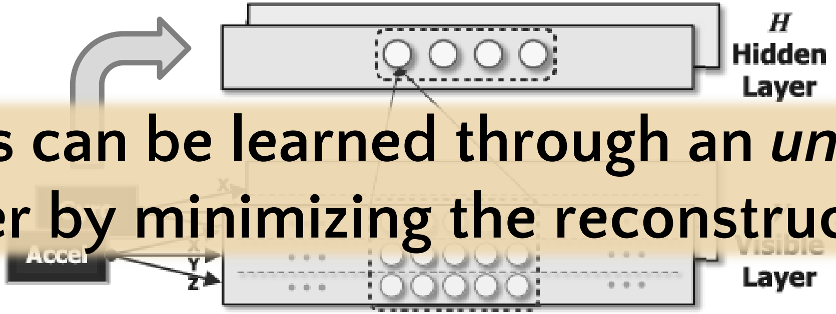
$$\mathbf{v} = \sum_k (W^k *_f \mathbf{h}_k) + a$$

CRBM Learns Motion Basis

Inference

$$P(\mathbf{h}_{k,j} = 1|I) = \sigma((\widetilde{W}^k *_v I)_j + b_k),$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Kernels can be learned through an *unsupervised* manner by minimizing the reconstruction error.

Reconstruction

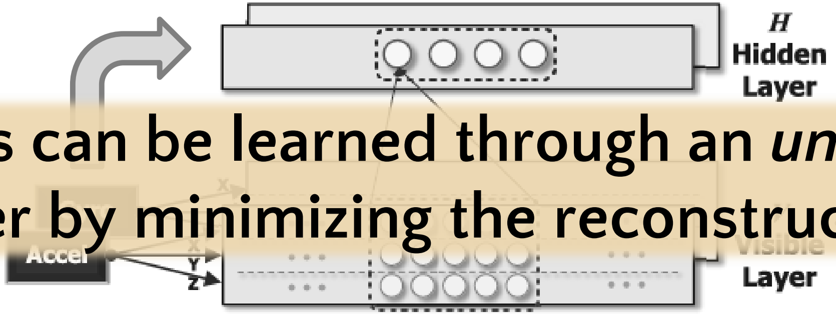
$$\mathbf{v} = \sum_k (W^k *_f \mathbf{h}_k) + a$$

CRBM Learns Motion Basis

Inference

$$P(\mathbf{h}_{k,j} = 1|I) = \sigma((\widetilde{W}^k *_v I)_j + b_k),$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



Kernels can be learned through an *unsupervised* manner by minimizing the reconstruction error.

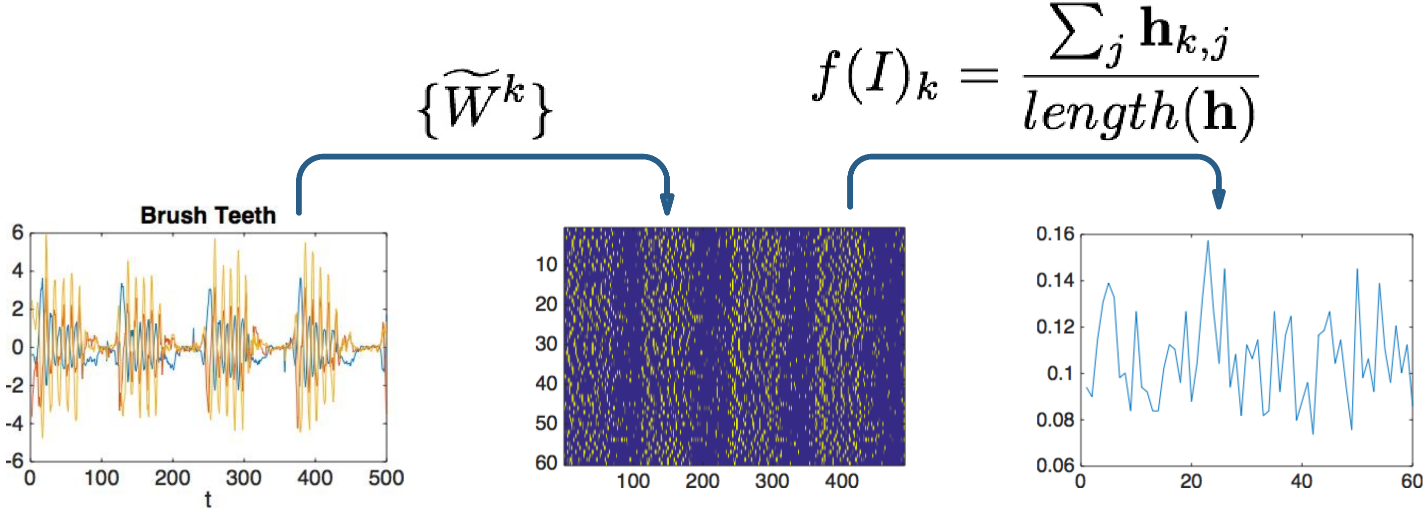
Reconstruction

$$\mathbf{v} = \sum_k (W^k *_f \mathbf{h}_k) + a$$

Basis

Embedding√

Semantic Descriptor Extraction



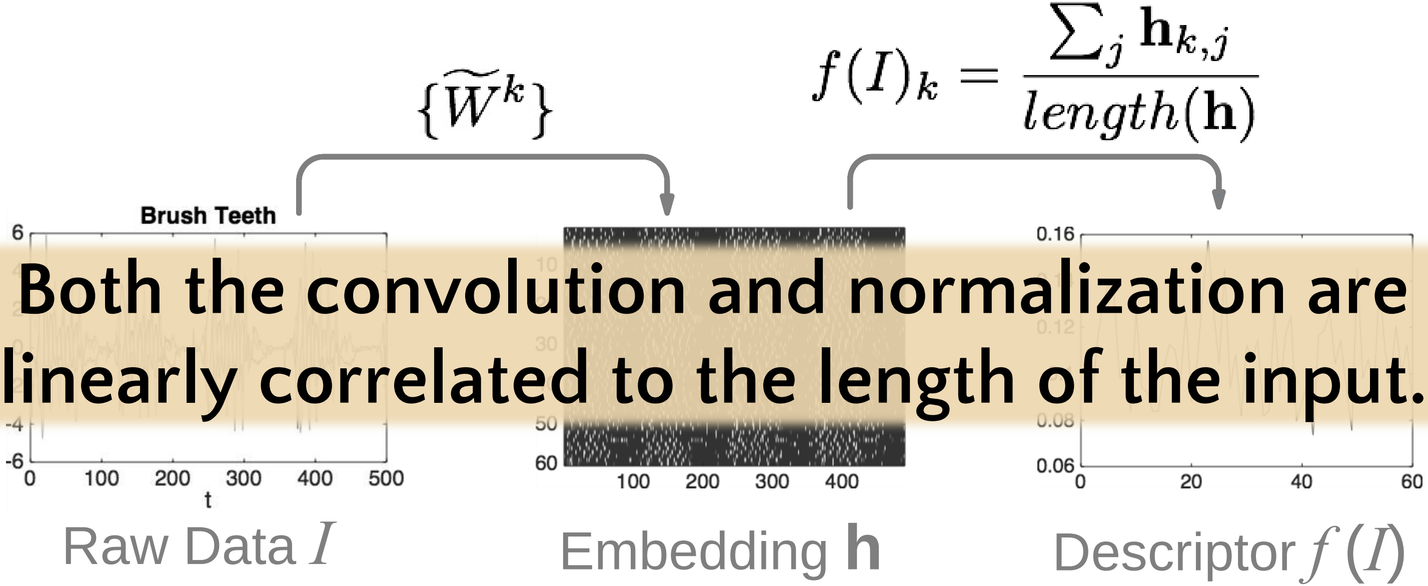
Raw Data I

Embedding $\mathbf{h}$

Descriptor $f(I)$

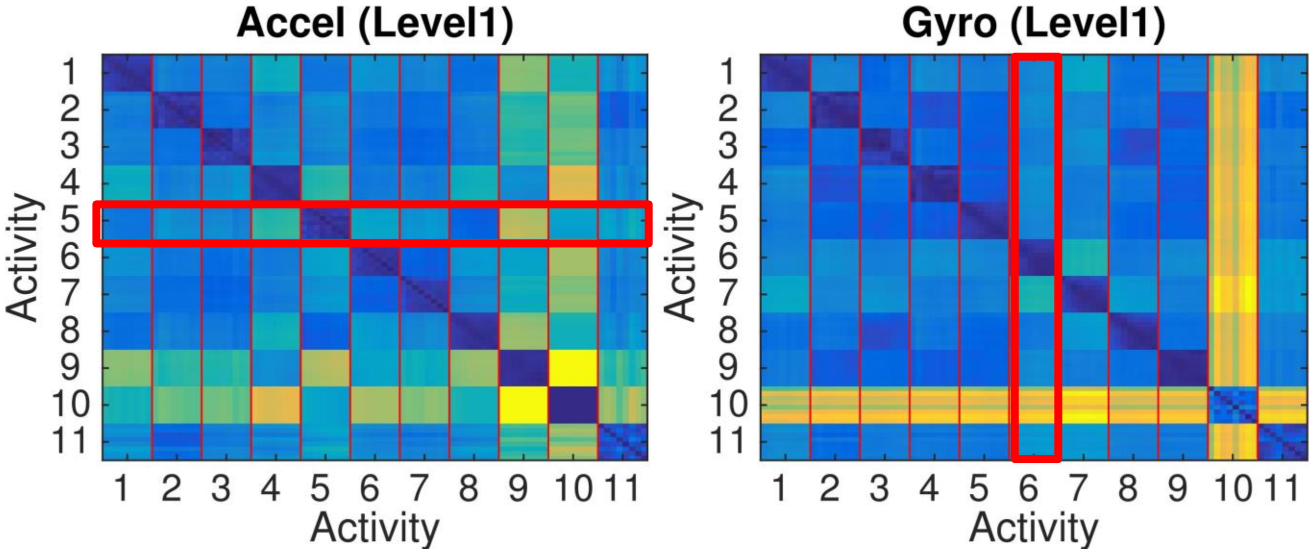
k=60

Semantic Descriptor Extraction



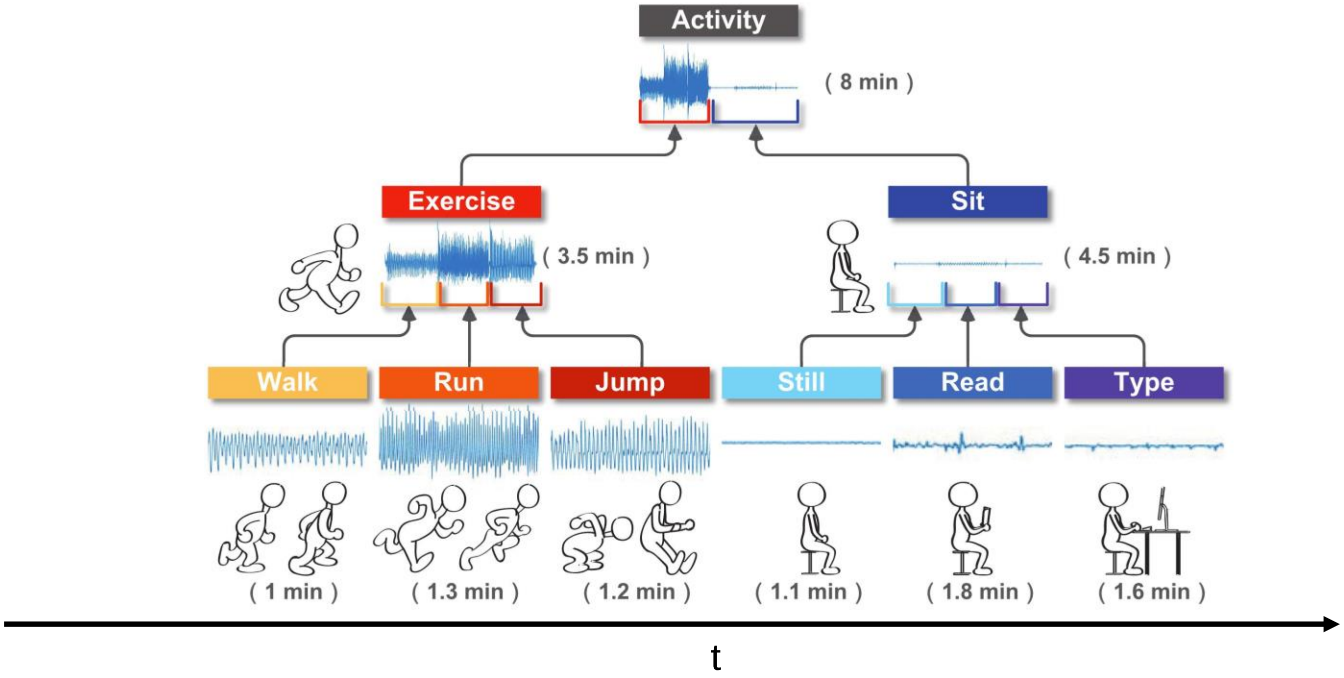
Both the convolution and normalization are linearly correlated to the length of the input.

Semantic Descriptor Extraction

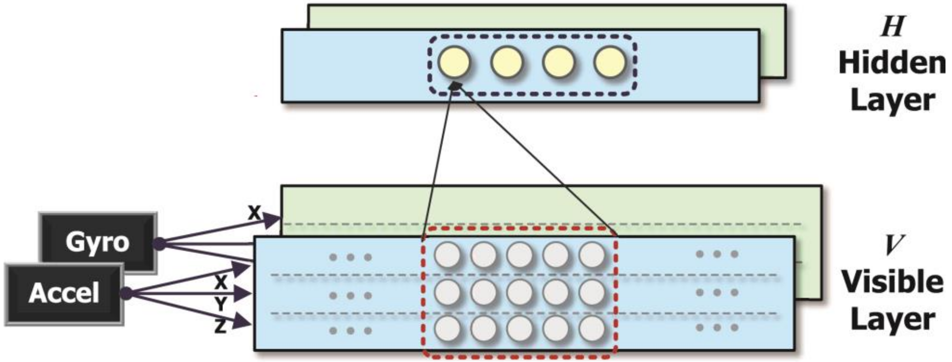


- Our descriptor helps to distinguish different activities.

Q2: How to address hierarchical semanteme?

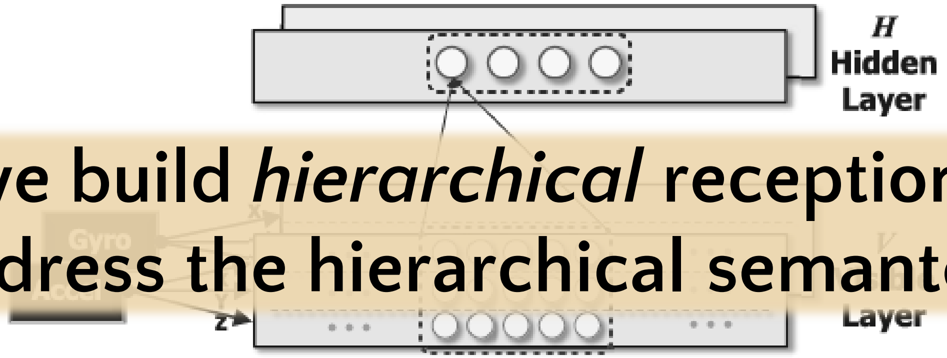


Inspiration: Reception Field



- The ***Reception Field*** refers to the kernel size. (3*5 in the figure)

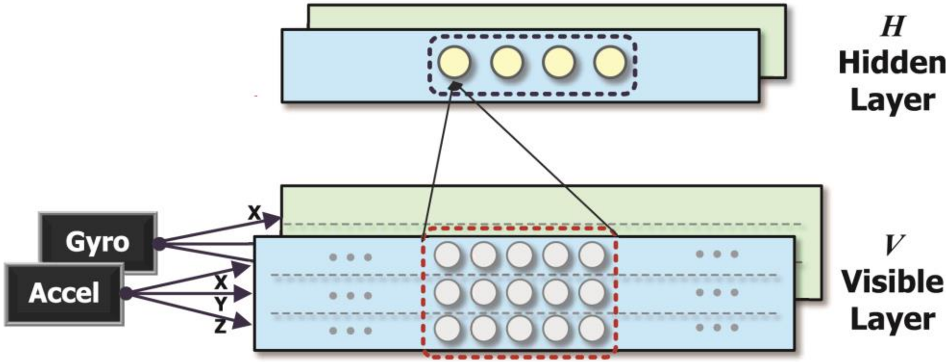
Inspiration: Reception Field



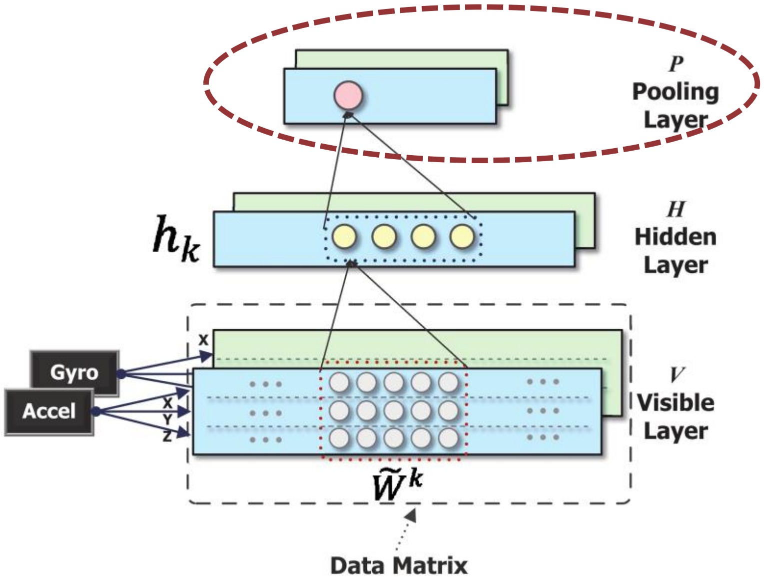
Can we build *hierarchical* reception field to address the hierarchical semanteme?

- The ***Reception Field*** refers to the kernel size. (3*5 in the figure)

Idea: Hierarchical Reception Field

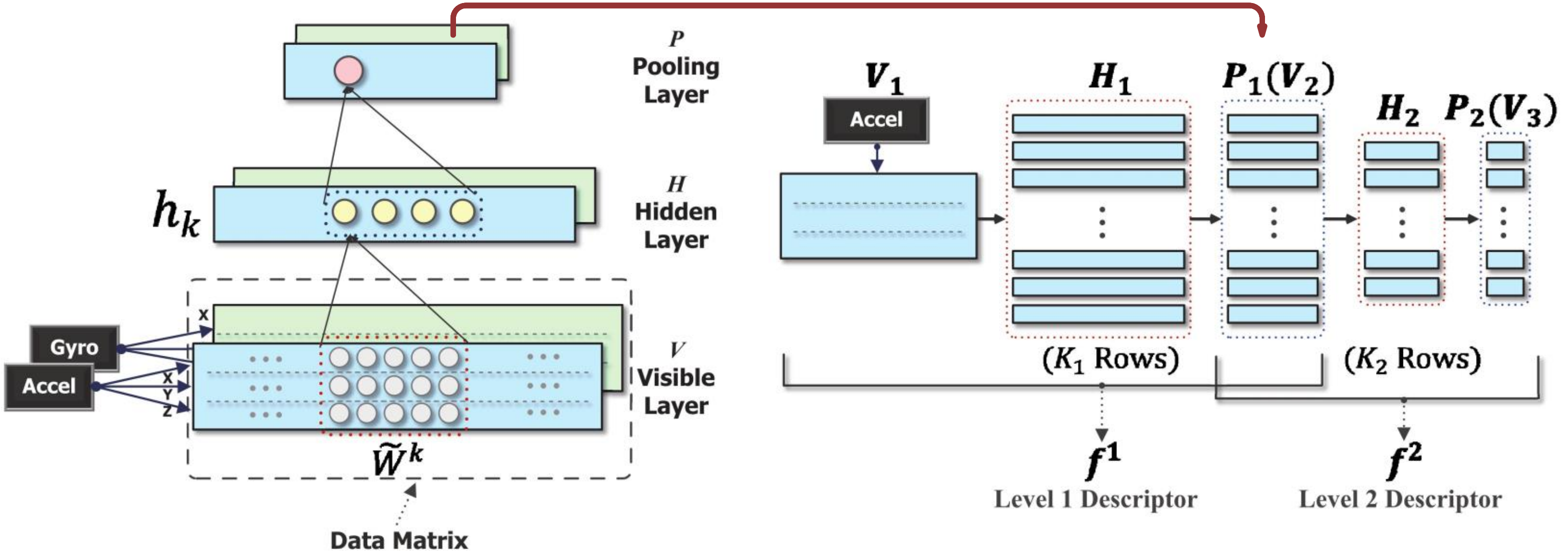


Idea: Hierarchical Reception Field



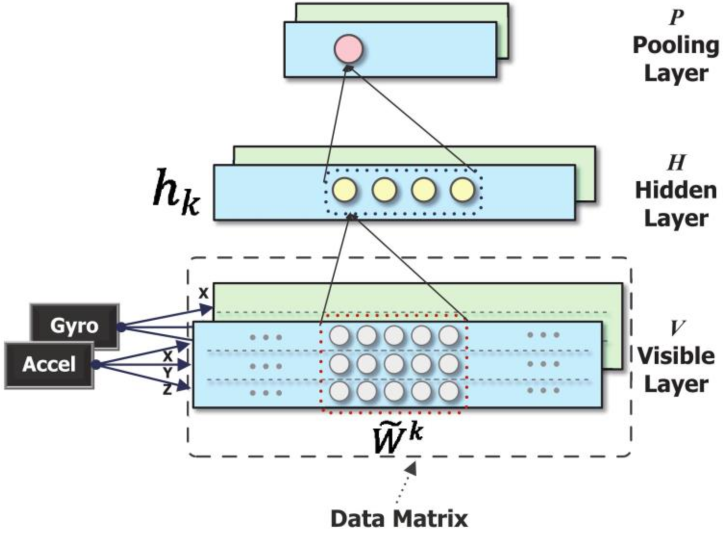
- 1. Add an ***Pooling Layer*** *P* to pool the output of *H*.

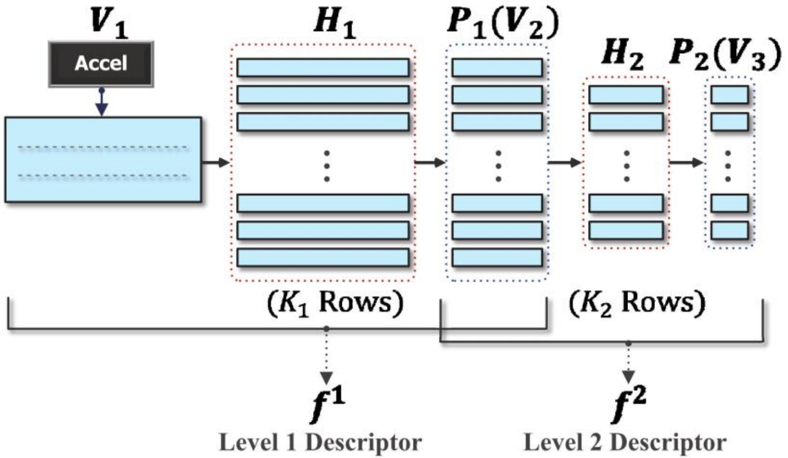
Idea: Hierarchical Reception Field



- 2. Stack multiple building blocks (feed V_2 with P_1).

Idea: Hierarchical Reception Field





💡 Kernels in *higher* level have *larger* reception field!

4.Semantic Based Activity Search

SBAS

- Retrieve the ***timespans*** of the same activity according to the ***activity*** performed by an querier in ***massive continuous*** mobile sensing data.

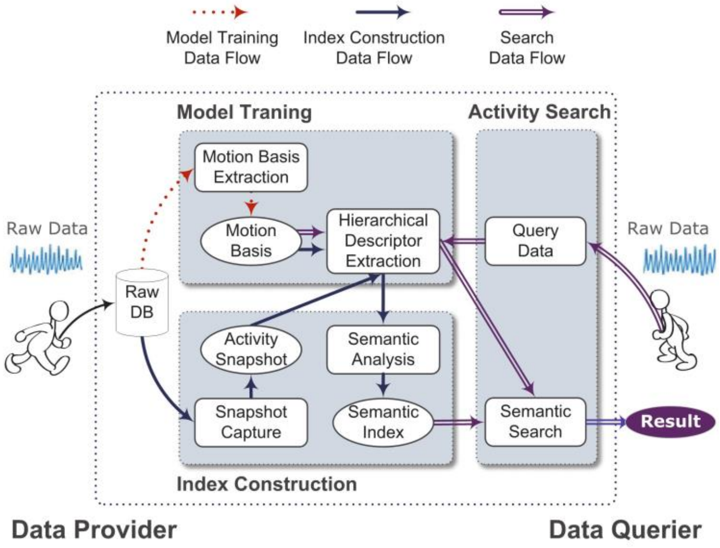
SBAS

- Retrieve the ***timespans*** of the same activity according to the ***activity*** performed by an querier in ***massive continuous*** mobile sensing data

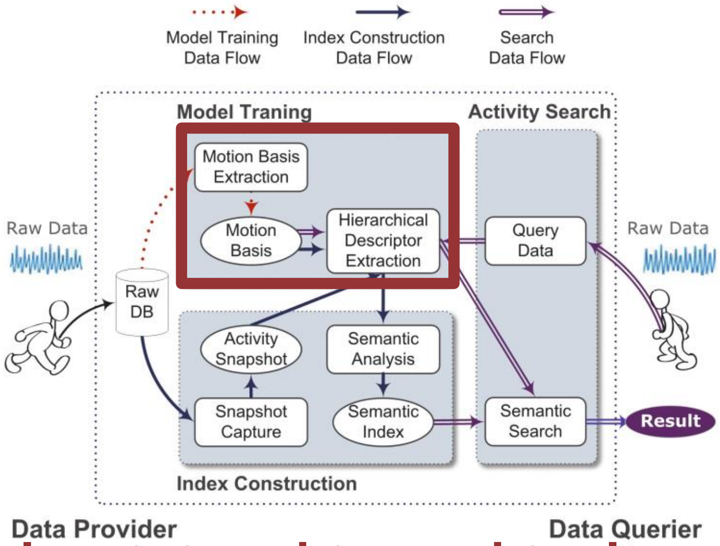
Different activities must get *separated*.

The search strategy must be *efficient*.

SBAS Architecture

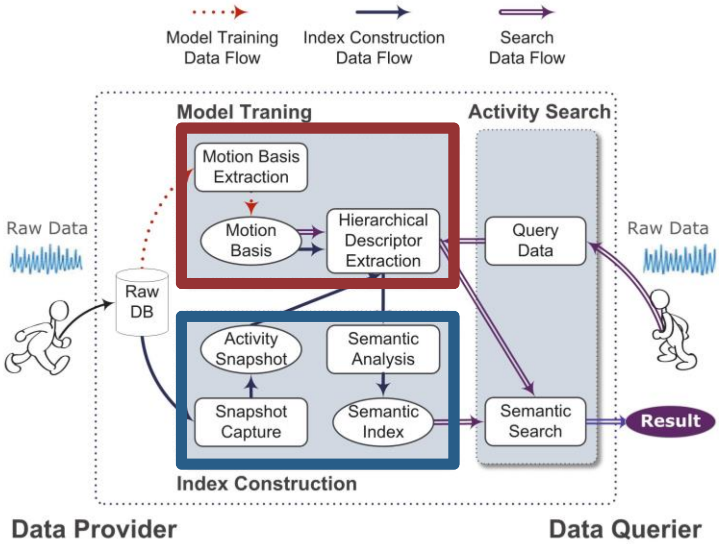


SBAS Architecture



- After model training, hierarchical motion basis is learned and descriptors can be extracted.

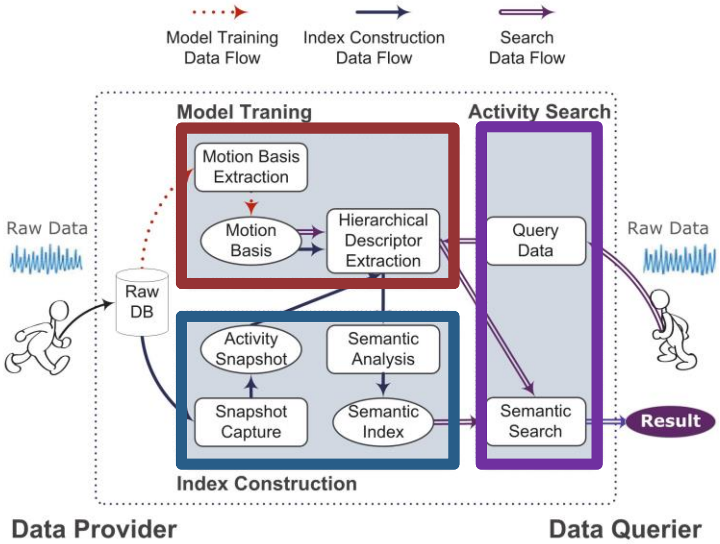
SBAS Architecture



- **Index Construction:**

1. Take activity snapshots using different timescale
2. Cluster the snapshots according to their descriptors

SBAS Architecture



- Search:

1. Perform *cluster search* in the index
2. *Merge* the timespans of the cluster search results

5. Implementation & Evaluation

Implementation

- **Android**

Sony Smartwatch3

- **Tizen**

Samsung Galaxy Gear

Client Side

Server Side

Model Training Server

- 4GHz i7 CPU
- Titan x-12G
- 32G Ram

SBAS Server

- 2.5GHz i7 CPU
- 16GB RAM

Architecture

Dataset

- **#1(controlled)**

8 people (M:7,F:1)
11 activities

- **#1(uncontrolled)**

8 people (M:7,F:1)
11 + x activities

- **#2(controlled)**

10 people (M:7,F:1)
7 activities

2.7GB(Over 320 hours)

323.9MB

Evaluation – Semantic Descriptor

Sensor	[14]	[15]	1-level	2-level	3-level
Accel	80.3	-	94.6	96.1	98.4
Gyro	71.8	-	82.1	82.9	91.4
Accel+Gyro	90.3	98.75	97.8	98.2	98.9

- For dataset#2, our 2-level hierarchical descriptor can provide comparable accuracy and the 3-level descriptor can provide even better performance.

*[14] M.Shoaib, S.Bosch, O.D.Incel, H.Scholten, andP.J.Havinga, "Fusion of smartphone motion sensors for physical activity recognition," Sensors, vol. 14, no. 6, pp. 10 146–10 176, 2014.

[15] W.Jiang and Z.Yin,"Human activity recognition using wearable sensors by deep convolutional neural networks," in Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, 2015, pp. 1307–1310.

Evaluation – Semantic Based Activity Search

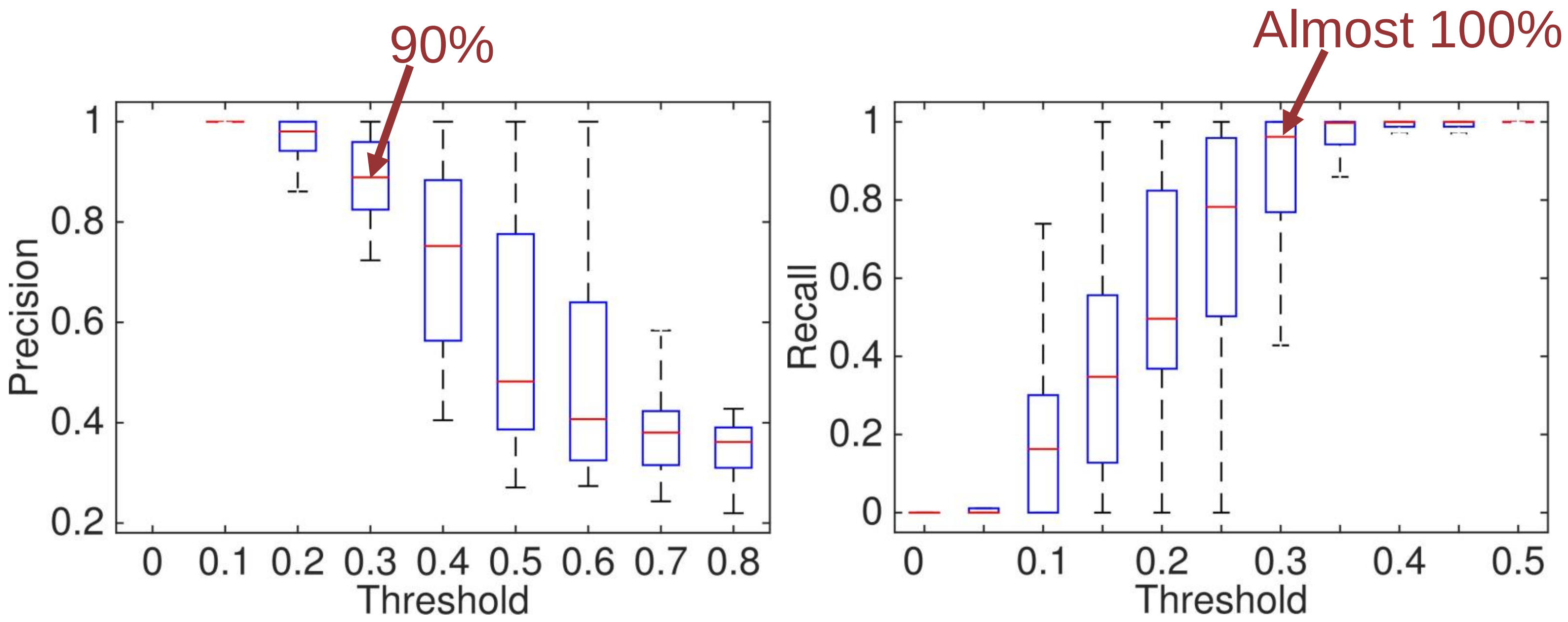
Three kinds of metrics are adopted:

- **Precision**
- ***Recall***
- ***Time Overhead***

**We adjust the search threshold to evaluate the precision and recall. Intuitively we have the tradeoff ,*

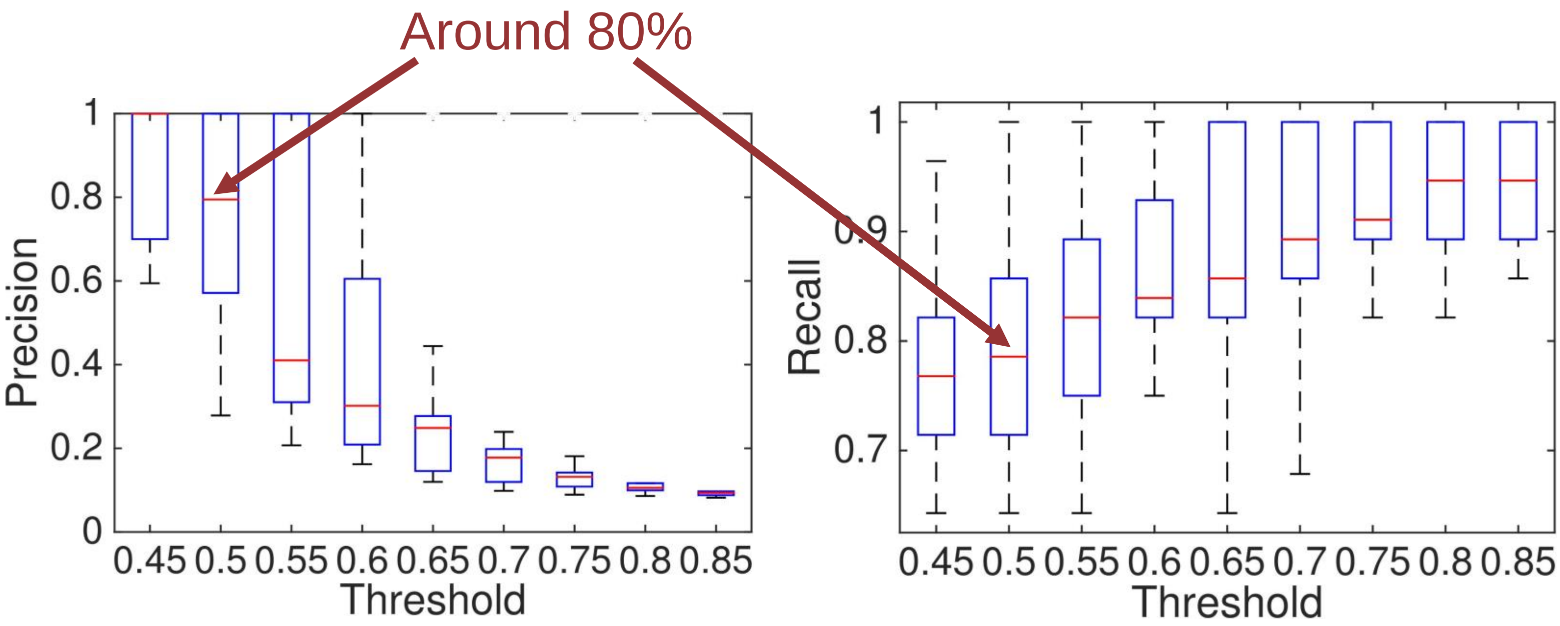
Similarity Threshold $\uparrow$ $\rightarrow$ *Precision* $\downarrow$ + *Recall* $\uparrow$

Evaluation – Semantic Based Activity Search



- For Dataset [#1](**controlled**), when the threshold is set to 0.3, an 90% precision and almost 100% recall can be achieved.

Evaluation – Semantic Based Activity Search



- For Dataset [#1](**uncontrolled**), the decline is caused by the complex human motion and mislabeled groundtruth in the *uncontrolled* environment.

Evaluation – Semantic Based Activity Search

Data Size	1min	10min	1h	1d	10d(>2Gb)
Indexing Time(s)	0.001	0.02	0.55	7.89	71.63
Search Time(s)	0.0008	0.002	0.052	0.28	8.83

- Keeping running ***Lasagna*** at backstage only leads to about *10%* additional power consumption.

6. Conclusion & Open Issues

Conclusion

Deep hierarchical understanding

- Motion basis is learned in an unsupervised manner.
- Hierarchical semantic descriptor is extract from different resolutions.

Semantic Based Activity Search

- Efficient SBAS can be achieved on COTS laptop.

Open Issues

- Database preprocessing
  - Activity Segmentation
  - Indexing
  - …
- More advanced searching strategies
  - Cross-modal SBAS
- Privacy issues
- …

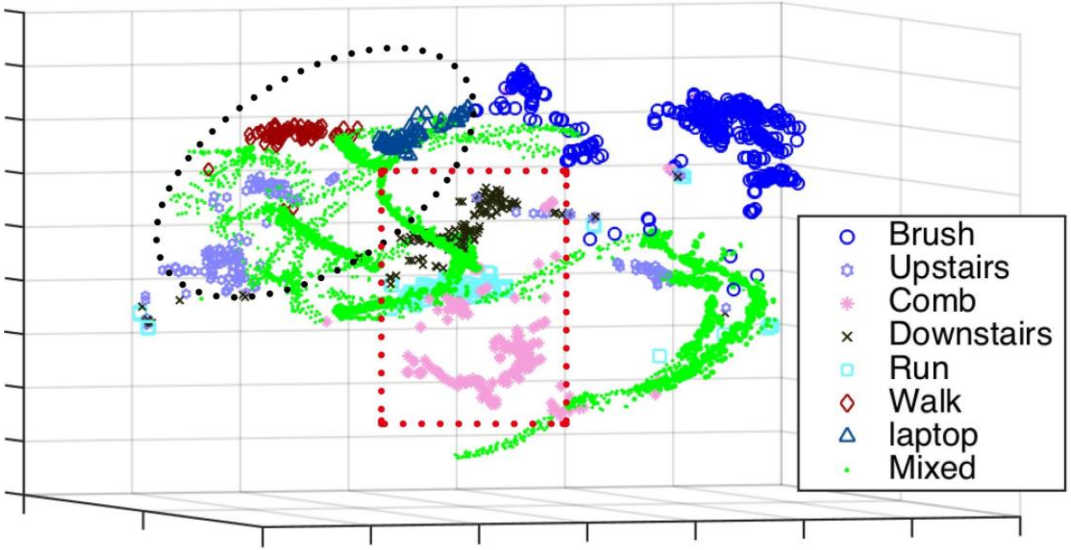
Thanks!

Any questions?

Feel free to contact me at

cihang@greenorbs.com

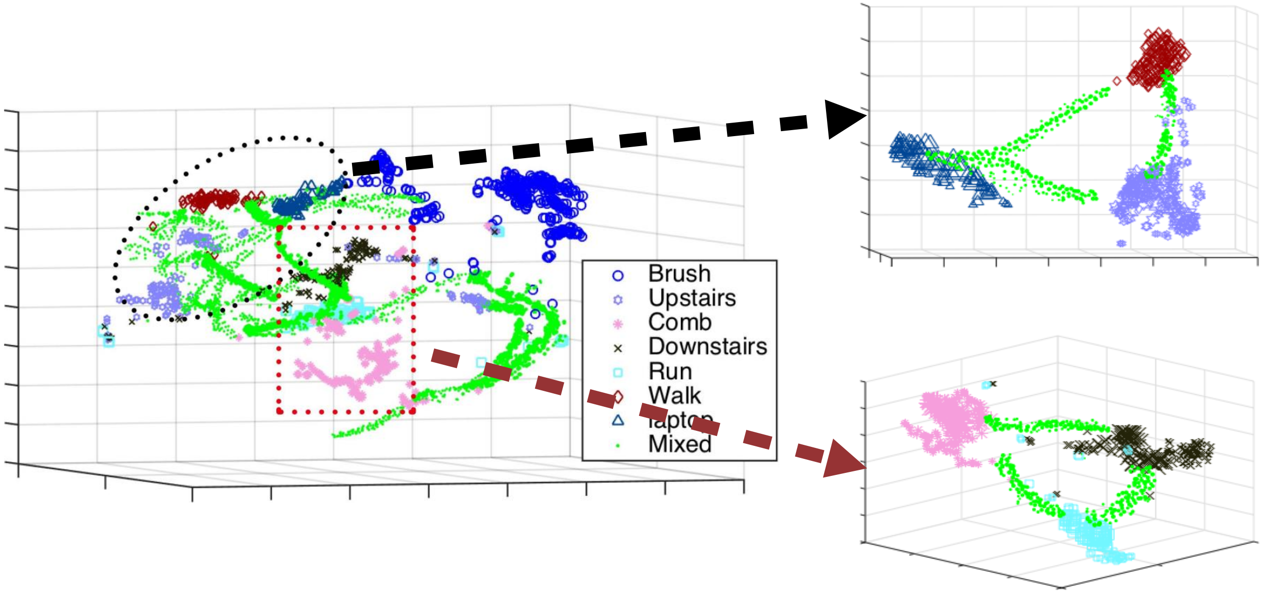
Hierarchical Semantic Descriptor



- Descriptors of a same activity **cluster** together.

** 2-level hierarchical descriptor with Euclidean distance as the similarity measure*

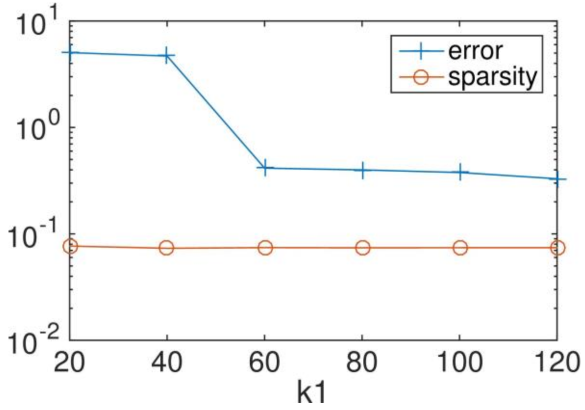
Hierarchical Semantic Descriptor

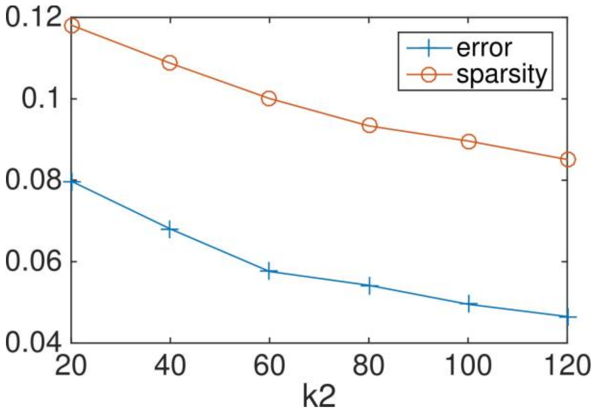


- Mixed activities *bridge* those “pure” activities.

** 2-level hierarchical descriptor with Euclidean distance as the similarity measure*

Evaluation – Kernel Number Selection

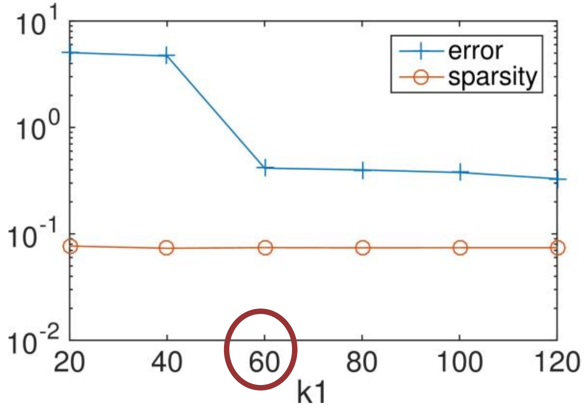


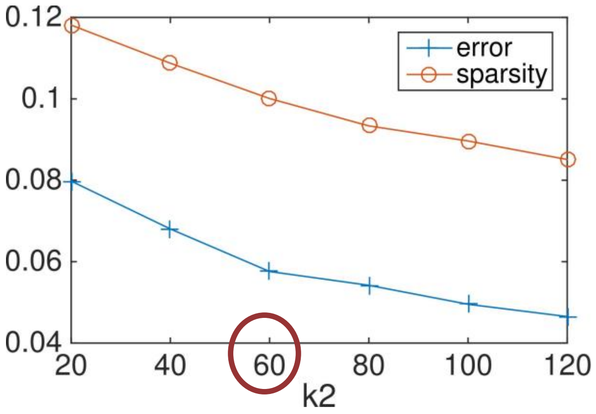


- A larger number of kernels helps to reduce error and sparsity.

**error*: |*input–reconstruction*|, *sparsity*: *mean*(*h*)

Evaluation – Kernel Number Selection

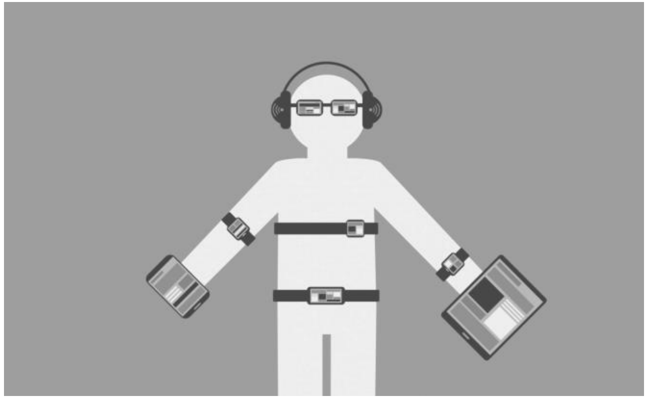




- A larger number of kernels will also bring extra cost for storage and computation.

**error*: |*input-reconstruction*|, *sparsity*: *mean*(*h*)

The Unsatisfying Smart Wearables



What wearables can do:

- Step counting
- Step counting
- …

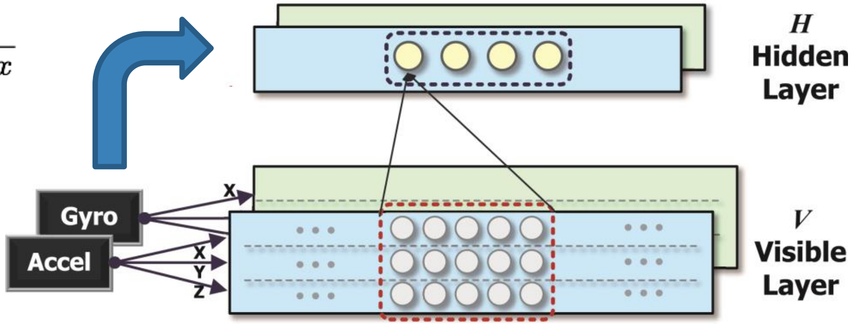
Is step counting the only thing that smart wearables can do?

CRBM Learns Motion Basis

Inference

$$P(\mathbf{h}_{k,j} = 1|I) = \sigma((\widetilde{W}^k *_v I)_j + b_k),$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



For example, with a kernel $\widetilde{W}^k$, 3*5 input units are mapped to 1 unit in the hidden layer.