Probabilistic multimodal map-matching with rich smartphone data

Jingmin Chen, Michel Bierlaire

Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne

July 16, 2012

Outline

Introduction

Objective: infer path and mode information of multimodal trips in a probabilistic manner from rich smartphone data.

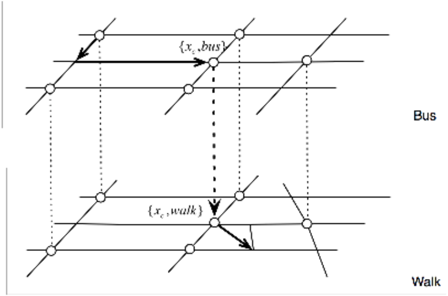
Motivations:

- Probabilistic method in order to account for errors in the data.
- Smartphone data provides rich mobility related information.
  - GPS can be used to detect the travel path.
  - Bluetooth and acceleration provide travel mode information.

Network representation

- Urban transport networks: walk, bike, car, bus, metro.
- A position $\mathbf{x} = (x, m)$ is characterized by horizontal coordinates x and transport mode m.
- Data source: OpenStreetMap.

Example: a multimodal path p in bus and walk networks.



Smartphone data collection

Nokia Data collection campaign:

- September 2009 - September 2011.
- 200 individuals in Geneva Lake area.
- Each individual is given with a Nokia N95 smartphone.
- A data collection APP constantly collect smartphone data and send it to a remote server.

Data used in this study:

- GPS, Bluetooth and Acceleration.
- Useful measurements are extracted from the raw data.

Smartphone measurements extraction

- GPS. 10 seconds interval. Coordinates $\widehat{x}$ with error indicators, speed $\widehat{v}$, heading $\widehat{h}$.
- Bluetooth (BT). 180 seconds interval. $\widehat{b}$ is equal to 1 if there is at least one BT device nearby, and 0 otherwise.
- Acceleration (ACCEL). 10 seconds ($40Hz$) recording with 120 seconds interval. A measurement $\widehat{a}$ is the mean of the accelerations in a 2-seconds time window (5 measurements for 10 seconds recording), unit $\frac{1}{280}m/s^2$.

Each measurement is associated with a time tag $\widehat{t}$.

Labelled data

For the calibrations of some models, we use measurements that are labelled with the true transport mode. The data is collected from 3 smartpthone users.

Number of measurements that are labelled:

	walk	bike	car	bus	metro	total
GPS	9350	11899	1142	1669	2069	26129
BT	non-PT:1826			PT: 869		2695
ACCEL	4501	11924	motor: 11801			28226

Measurements sequence from a trip

$\widehat{y}_{1:T} = (\widehat{y}_1, \ldots, \widehat{y}_T)$

- A sequence of measurements recorded during a continuous travel without intermediate stops for activities (usually a trip or a part of a trip).
- The measurements are chronologically ordered.
- It is composed of 3 subsequences: the GPS, the Bluetooth and the acceleration. E.g, the GPS measurements sequence, $(\widehat{g}_1, \widehat{g}_2, \ldots, \widehat{g}_I)$.

Probabilistic measurement model

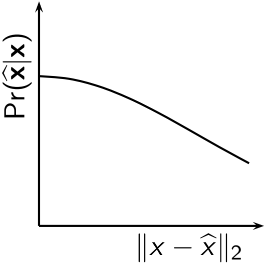
$\Pr(\widehat{y}_{1:T} | t_{1:T}, p)$ calculates the likelihood of observing all the smartphone measurements $\widehat{y}_{1:T}$ on a multimodal path p at time $t_{1:T}$.

Components:

1. Sensor measurement models: captures the data generation processes.
2. Travel model: captures the dynamics in travel.
3. Integration: integrate them in a unified framework.

GPS measurement model

- The errors in longitudinal and latitudinal directions are independently normal distributed.
- Then, the horizontal error e is Rayleigh distributed.
- The variance $\widehat{\sigma}^2$ is estimated from the GPS and network data.



$$\Pr(\widehat{\mathbf{x}}|\mathbf{x}) = \Pr(\widehat{x}|x) = \Pr(e > \|\widehat{x} - x\|_2) = \exp\left(-\frac{\|x - \widehat{x}\|_2^2}{2\widehat{\sigma}^2}\right)$$

Bluetooth measurement model

Assumption: Bluetooth only distinguishes between PT and non-PT.

$$\Pr(\widehat{b}|\mathbf{x}) = \Pr(\widehat{b}|m) = \begin{cases} \Pr(\widehat{b}|m \in \mathrm{PT}) & \text{if } m \text{ is PT} \\ \Pr(\widehat{b}|m \notin \mathrm{PT}) & \text{if } m \text{ is non-PT} \end{cases}$$

where $\mathrm{PT} = \{\mathrm{bus}, \mathrm{metro}\}$.

The model is estimated from the labelled data:

$\Pr(\widehat{b}|m)$	$\widehat{b} = 0$	$\widehat{b} = 1$
$m \in \mathrm{PT}$	0.19	0.81
$m \notin \mathrm{PT}$	0.60	0.40

There is a higher chance to observe a Bluetooth device in public transport environment.

Acceleration measurement model

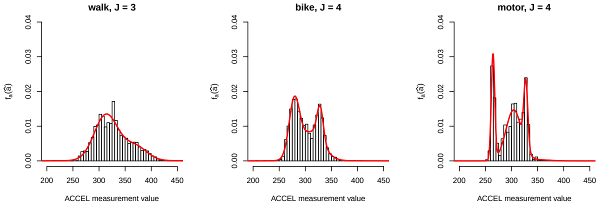
- Assumption: acceleration only distinguishes among: walk, bike, motor modes.
- We calibrate a pdf using a mixture of normal for each of them. Normal mixture is usually used to estimate distributions of heterogenous data (e.g. GPS speed data[1]).

$$\Pr(\widehat{a}|x) = \Pr(\widehat{a}|m) = f_a(\widehat{a}|m) = \sum_{j=1}^{J_m} w_{mj}\phi(\mu_{mj}, \sigma_{mj}^2). \tag{1}$$

1. Park B-J, Zhang Y, Lord D. Bayesian mixture modeling approach to account for heterogeneity in speed data. Transportation Research Part B: Methodological. 2010;44(5):662-673.

Acceleration measurement model

The pdf's are estimated from the labelled data.



- *Walk*: higher chance to observe a high acceleration.
- *Bike*: stable with a peak near gravity (280).
- *Motor*: a peak lower than the gravity, which depicts vertical movements caused by the road condition.

Travel model

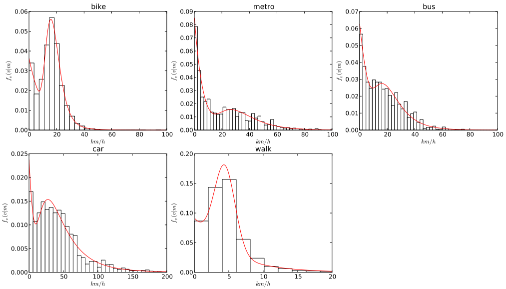
$$\Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_{k-1}, t_k, p)$$

predicts the position $\mathbf{x}_k = (x_k, m_k)$ at time t_k, given that the state at time t_{k-1} is $\mathbf{x}_{k-1} = (x_{k-1}, m_{k-1})$, and the smartphone user is traveling along path p.

We use GPS speed data to derive and calibrate the travel model.

The speed distribution $f_v(\widehat{v}|m)$ for each mode m

Mixture of a normal/lognormal (walk/others) and a negative exponential. Negative exponential captures stops. Normal/lognormal captures travel at regular speed.



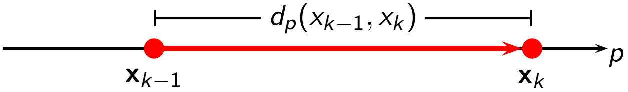
Remarks:

- Calibrated from labelled data.
- Normal fits better for walk.
- Lognormal fits better for others.

Derivation of the travel model

Assumption: at most one mode change between $\mathbf{x}_{k-1}$ and $\mathbf{x}_k$.

- Case 1: No mode change between $\mathbf{x}_{k-1}$ and $\mathbf{x}_k$.

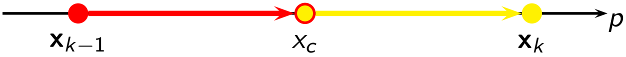


Assumption: the travel speed follows the speed distribution of transport mode m_k,

$$\Pr(x_k | x_{k-1}, t_{k-1}, t_k, p) = f_v\left(\frac{d_p(x_{k-1}, x_k)}{t_k - t_{k-1}} | m_k\right)$$

Derivation of the travel model

- Case 2: One mode change at x_c between $\mathbf{x}_{k-1}$ and $\mathbf{x}_k$. The mode change time t_c is unknown.



$$\Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_{k-1}, t_k, p) = \int_{t_c=t_{k-1}}^{t_k} \Pr(t_c|\mathbf{x}_{k-1}, t_{k-1}, p) \Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_c, t_k, p) dt_c$$

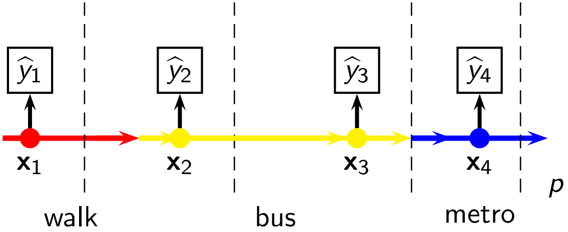
Assumption: each unimodal travel segment follows the speed distribution of the corresponding transport mode.

  - $\Pr(t_c|\mathbf{x}_{k-1}, t_{k-1}, p) = f_v(\frac{d_p(x_{k-1}, x_c)}{t_c - t_{k-1}}|m_{k-1})$,
  - $\Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_c, t_k, p) = f_v(\frac{d_p(x_c, x_k)}{t_k - t_c}|m_k)$.

Derivation of the smartphone measurement model.

Decomposition of the measurement model:

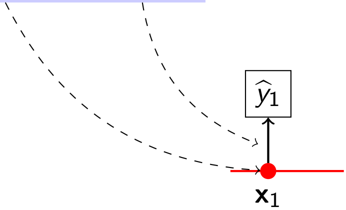
$$\Pr(\widehat{y}_{1:T}|t_{1:T}, p) \quad = \Pr(\widehat{y}_1|t_1, p) \prod_{k=2}^{T} \Pr(\widehat{y}_k|\widehat{y}_{1:k-1}, t_{1:k}, p),$$



The states $\mathbf{x}_{1:T}$ are latent

First iteration

$$\Pr(\widehat{y}_1|t_1, p) = \int_{\mathbf{x}_1 \in p} \Pr(\mathbf{x}_1|t_1, p) \;\; \Pr(\widehat{y}_1|\mathbf{x}_1) \, d\mathbf{x}_1$$



Two components:

1. A prior probability .
2. A sensor measurement model.

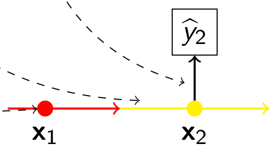
Each subsequent iteration

$$\Pr(\widehat{y}_k|\widehat{y}_{1:k-1}, t_{1:k}, p) = \Pr(\widehat{y}_k|\widehat{y}_{k-1}, t_k, p) = \int_{\mathbf{x}_k \in p} \Pr(\widehat{y}_k|\mathbf{x}_k)\Pr(\mathbf{x}_k|\widehat{y}_{k-1}, t_k, p) d\mathbf{x}_k$$

$$= \int_{\mathbf{x}_k \in p}\int_{\mathbf{x}_{k-1} \in p} \Pr(\mathbf{x}_{k-1}|\widehat{y}_{k-1}, p)\ \Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_{k-1}, t_k, p)\ \Pr(\widehat{y}_k|\mathbf{x}_k)\, d\mathbf{x}_{k-1} d\mathbf{x}_k$$

Three components:

1. A posterior probability from the last iteration.
2. A travel model.
3. A sensor measurement model.



Candidate path generation

Given a set of candidate paths P, we can infer how likely $p \in P$ is the true path:

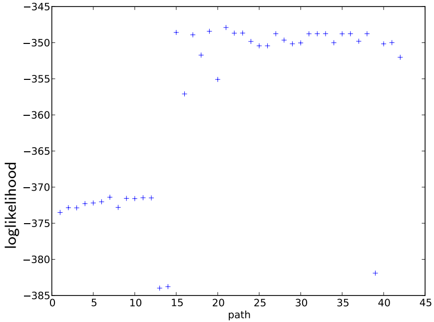
$$q(p|\widehat{y}_{1:T}) = \frac{\Pr(\widehat{y}_{1:T}|t_{1:T}, p)\Pr(p)}{\sum_{p' \in P} \Pr(\widehat{y}_{1:T}|t_{1:T}, p')\Pr(p')}, \tag{2}$$

where $\Pr(p)$ is a prior probability. We propose an algorithm to generate P:

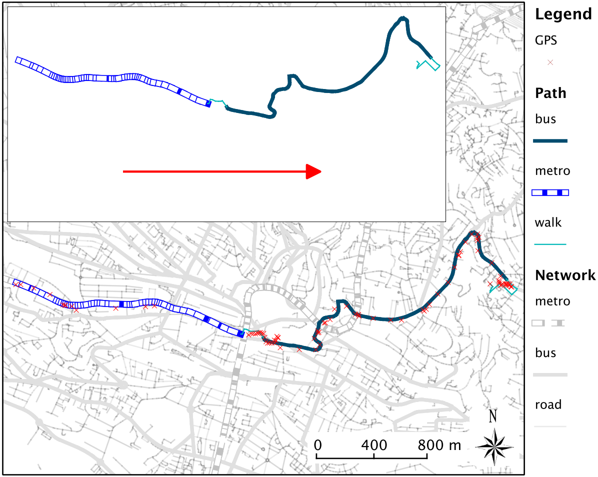
- The algorithm builds the physical path and the transport modes simultaneously.
- Smartphone data recorded in a multimodal trip are not required to be preprocessed into several unimodal segments.
- Transport networks also contribute to the inference of the transport mode, especially in differentiating PT and non-PT modes.

Result illustration: a multimodal trip

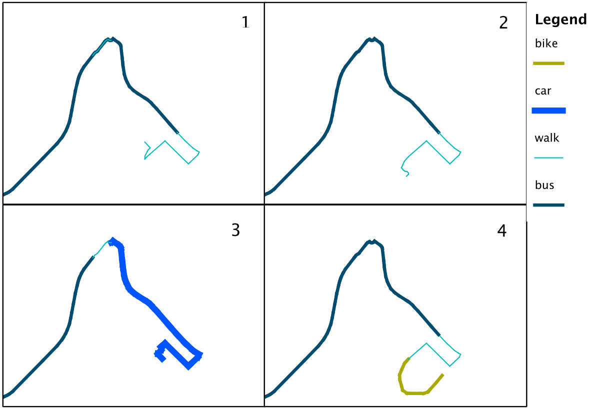
A multimodal trip: metro→walk→bus→walk (20 minutes).
Input:91 GPS, 8 BT, and 395 ACCEL.
Output: 43 multimodal paths with measurement loglikelihoods.



Result illustration: the most likely path



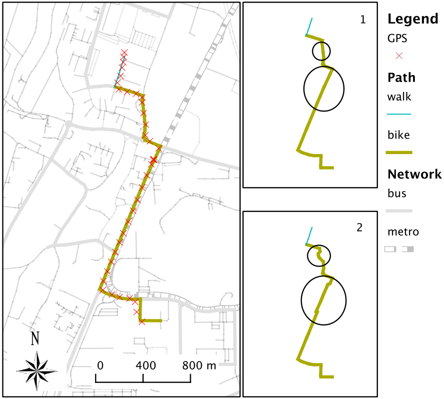
Result illustration: uncertainty in the trip end



loglikelihoods: -347.9, -348.7, -384.0, -381.9.

Result illustration: a bike trip

- A trip with bike as the main mode and walk at the end.
- 33 paths generated.
- Likelihood for two examples: -117.7, -118.0.



Performance analysis on transport mode inference

Data: 36 data sequences that are known to have one single mode.
$S(P', P) \in [0, 1]$ measures the similarity between two sets of paths.
1 indicates complete overlap, 0 indicates no overlap at all.

Input		Output	
data	known mode	path set	S
GPS	Yes	P^0	$S^0 = S(P_0, P_0)$
GPS	No	P^1	$S^1 = S(P_1, P_0)$
GPS, BT	No	P^2	$S^2 = S(P_2, P_0)$
GPS, BT, ACCEL	No	P^3	$S^3 = S(P_3, P_0)$

P^0 has the correct mode, hence it is served as the benchmark.
S^0 measures the uncertainty of P^0, the result with known mode.

Performance analysis on transport mode inference

Some examples

id	mode	time	GPS	BT	ACCEL	S^0	S^1	S^2	S^3
3	bus	234	24	1	11	0.96	0.64	0.65	0.93
9	car	229	20	0	23	0.97	0.95	-	0.96
16	bike	369	38	1	23	0.83	0.76	0.77	0.76
20	metro	560	34	1	23	0.99	0.77	0.82	0.85
34	walk	359	27	1	0	0.80	0.75	0.73	-

- In general, $S^0 > S^3 > S^2 > S^1$.
- PT results have higher accuracy, thanks to the lower density of PT networks.
- Average: S^2 (0.888) $>$ S^1 (0.858), in 12 cases where BT data is available.
- Average: S^3 (0.826) $>$ S^1 (0.751), in 22 cases where ACCEL data is available.

Conclusion remarks

- A flexible modeling framework is proposed:
  - The prior probability $Pr(p)$ is flexible.
  - Can integrate other types of data by defining the corresponding sensor measurement models.
  - Can integrate other networks, e.g. train network.
- Results analysis:
  - Results are consistent with the reality and the assumptions.
  - Capable of dealing with mode changes.
  - Good performance in identifying the transport modes.
  - Apart from the most useful GPS data, BT and ACCEL also contribute in identifying the transport mode.