Summarization of Egocentric Moving Videos for Generating Walking Route Guidance

Masaya Okamoto and Keiji Yanai
Department of Informatics,
The University of Electo-Communications,
Tokyo, Japan

Background

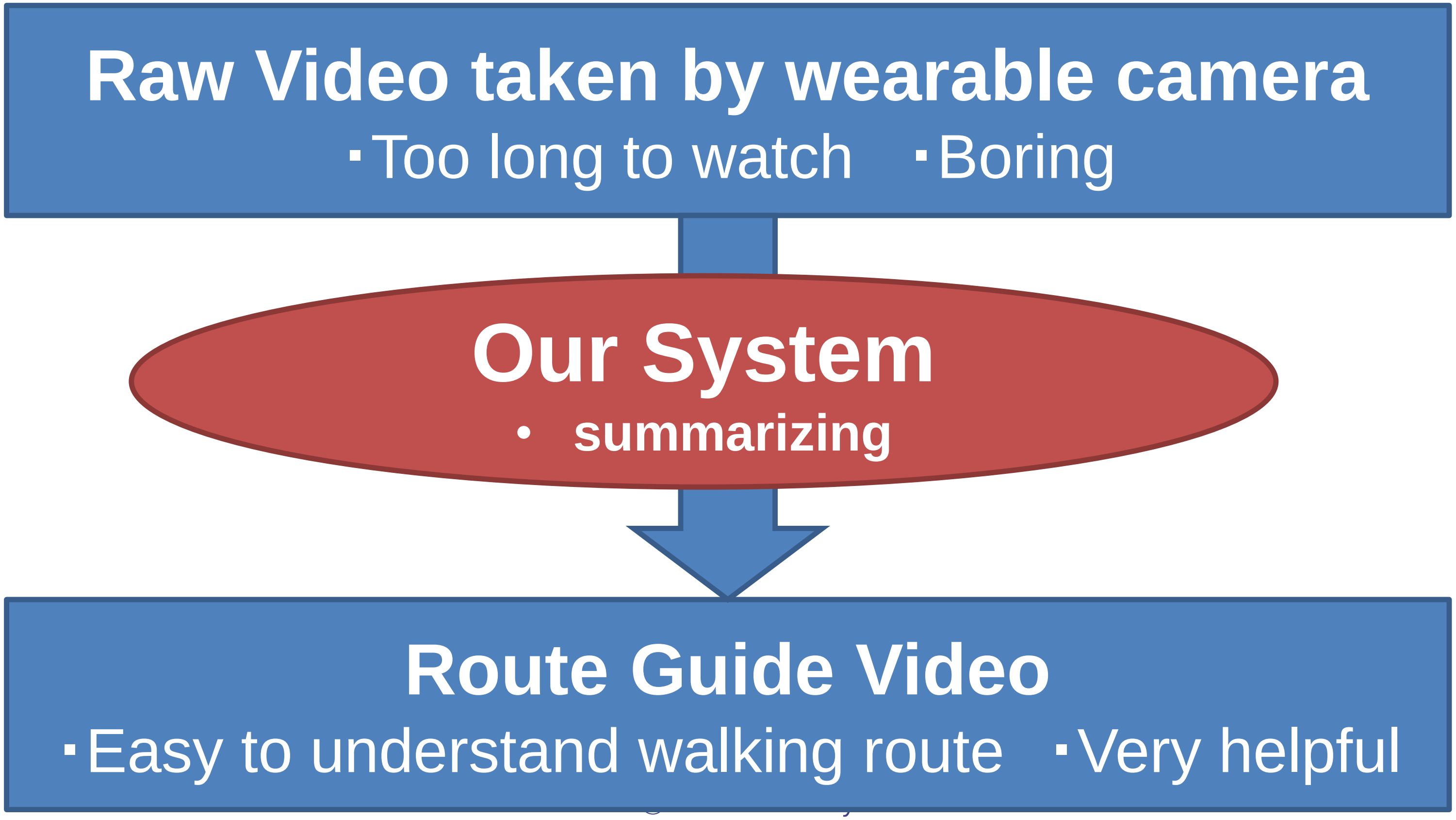
- Spread of wearable camera
  - Easy to take egocentric video

- New application of egocentric videos

Automatic generation of route guidance video

Objective

Generate walking route guidance
by summarizing egocentric moving videos



Raw Video taken by wearable camera

･Too long to watch ･Boring

Our System

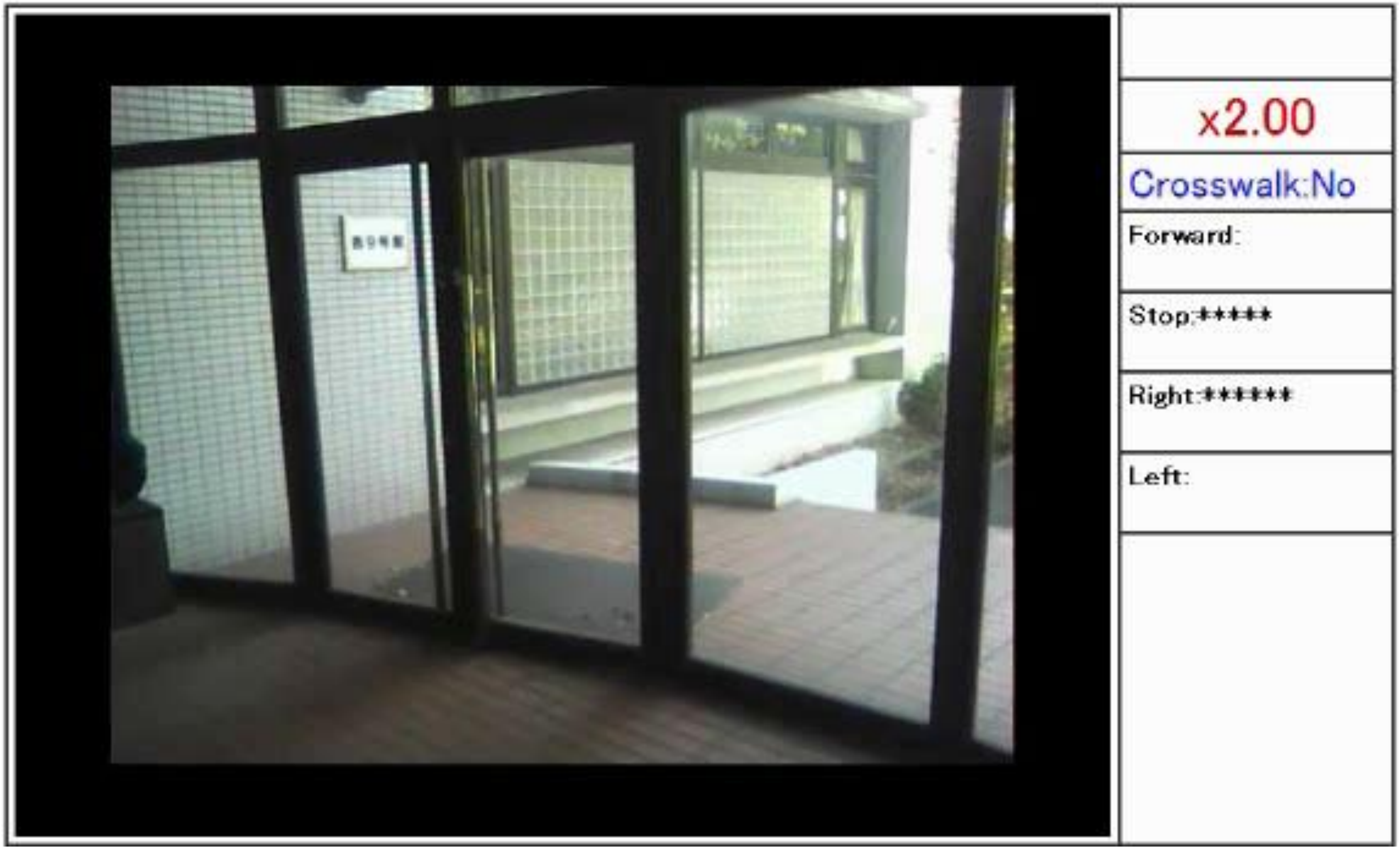
- **summarizing**

Route Guide Video

･Easy to understand walking route ･Very helpful

Demo (Raw Video)

Demo (Result Video)



x2.00

Crosswalk:No

Forward:

Stop:*****

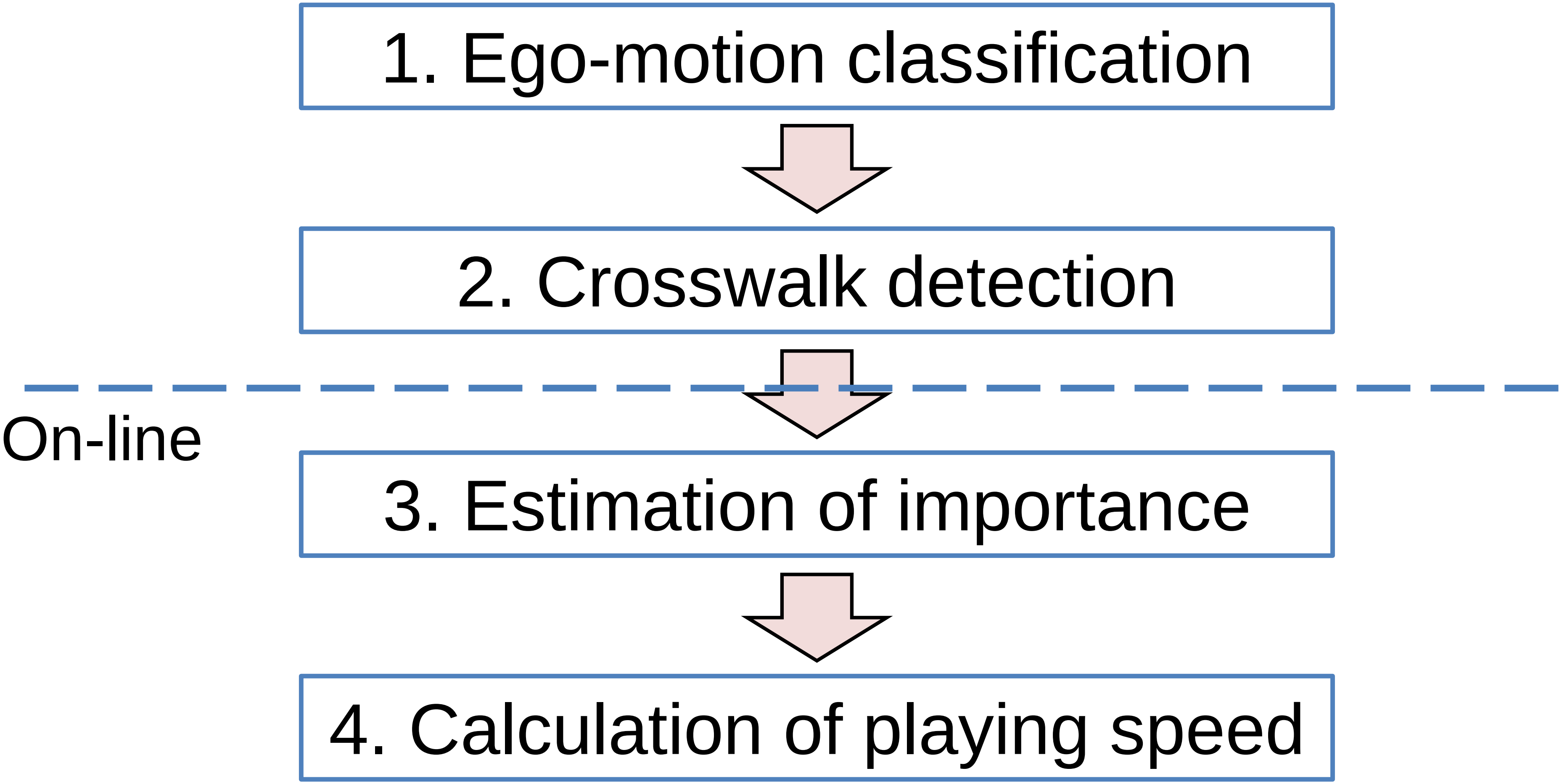
Right:******

Left:

Related Work: Summarization egocentric video

- Tancharoen et al. [ACM SIGMM 2005]
  - Cues: GPS and other sensors
  - Target: Life-log video (everyday life)
  - Output: Set of important frames
- Lee et al. [CVPR 2012]
  - Cues: Visual features
  - Target & Output: same as the above
- Ours
  - Cues: Visual information
  - Target: Walking video
  - Output: Summarized video

Overview



1. Ego-motion classification

2. Crosswalk detection

On-line

3. Estimation of importance

4. Calculation of playing speed

1. Ego-motion Classification

- Classify video sections into four classes

① Moving forward

② Stopping

③ Turning right

④ Turning left

Video section is four seconds long

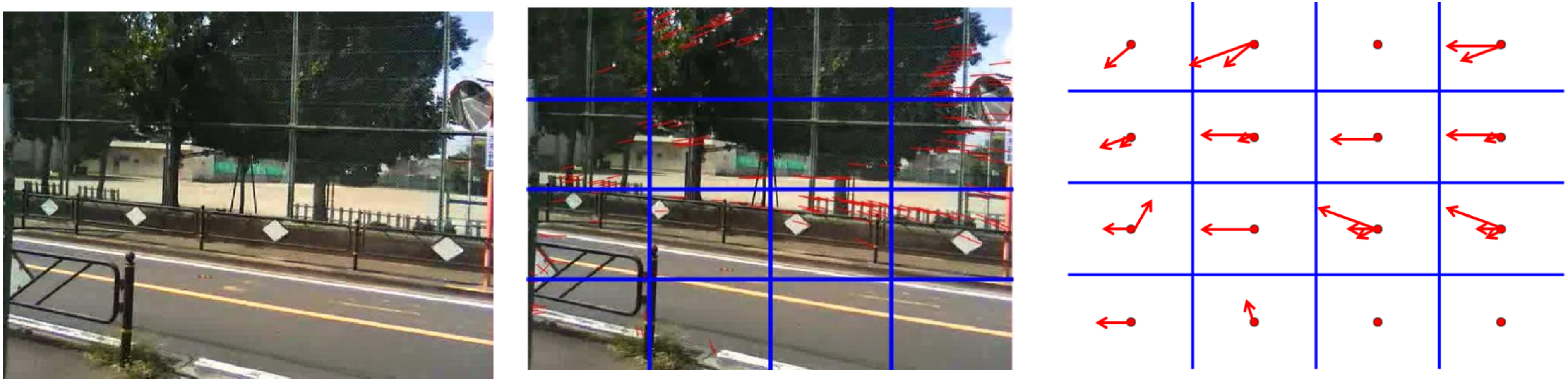
1. Ego-motion Classification

- Extract 48 frames from one video section
- Calculate a feature vector

① Compute optical flows for 47 intervals

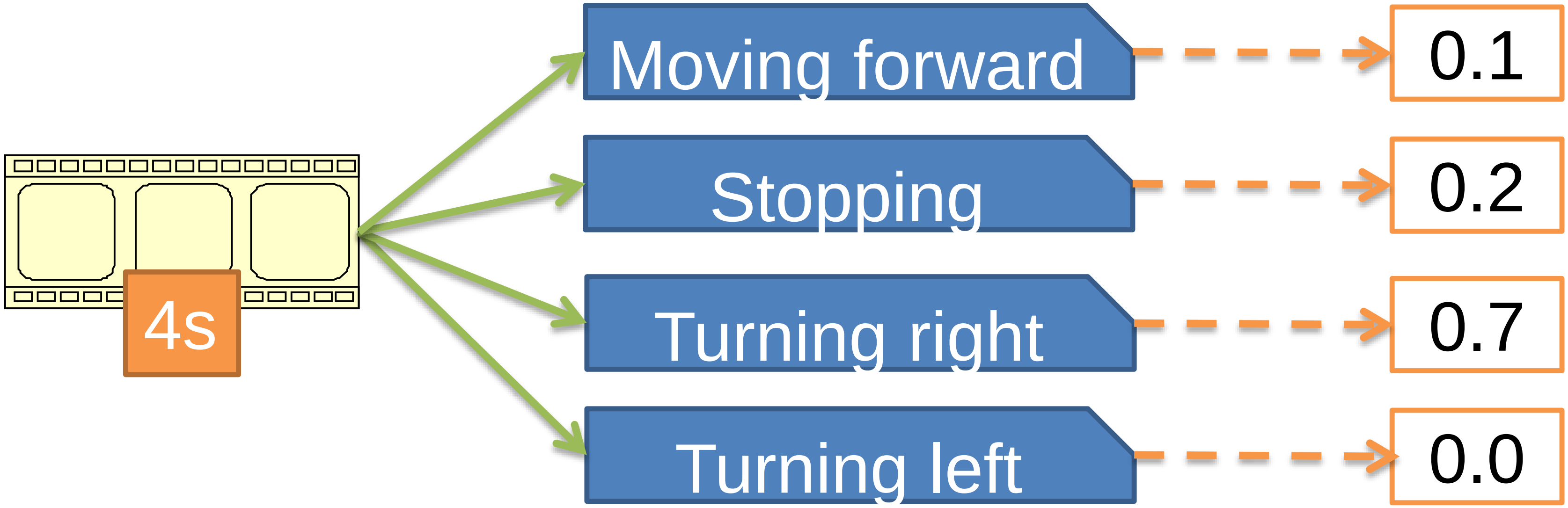
② Build 18-bin directional histograms for 4x4 grids

③ Normalize them within a video section



1. Ego-motion Classification

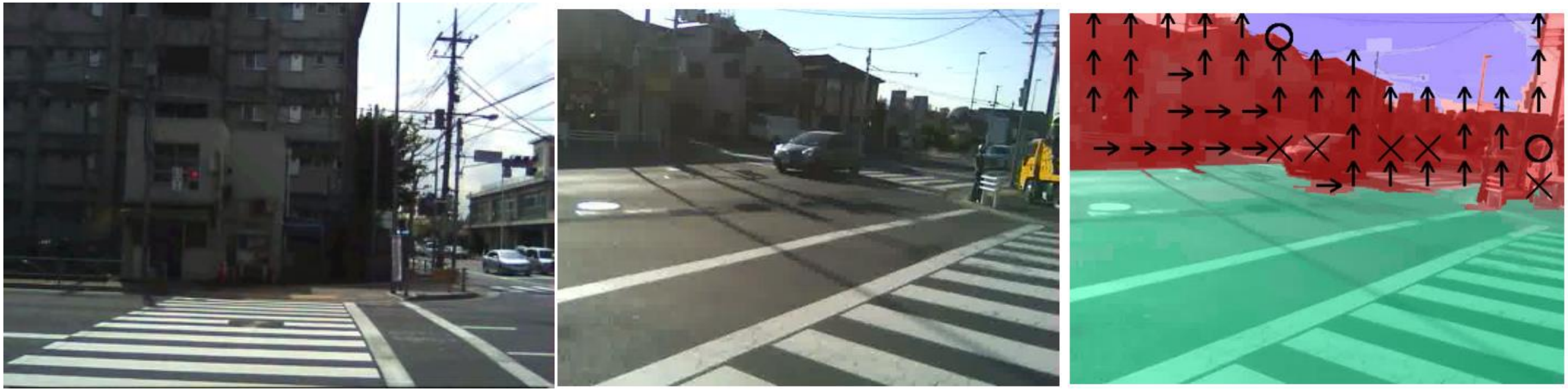
- Train 4 SVM classifiers in one-vs-all
  - Prepare hand-labeled training data
- Use pseudo-probability values
  - To estimate section importance



2. Crosswalk Detection

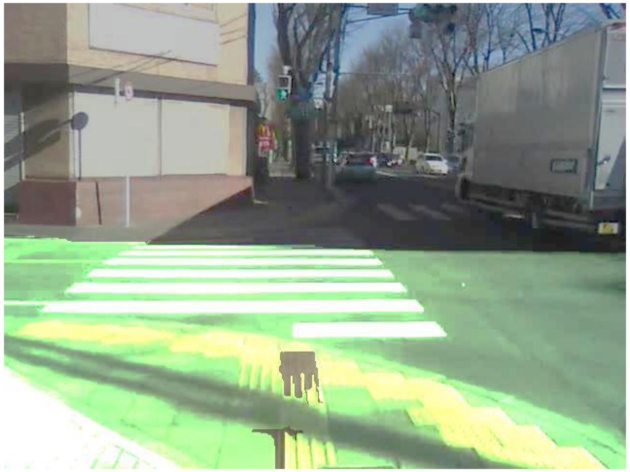
Crosswalk is important and remarkable cue

- Extract three frames every second
- Estimate ground regions
    - – Use Geometric context (Hoiem et al. [IJCV Vol.75 2007])



2. Crosswalk Detection

- Extract SIFT feature from ground regions
- Make BoF vector with extracted SIFT
- Use non-linear SVM as classifier
  - – about 240 learning frames



3. Estimation of Importance

- Expression of estimation importance

$$S_i = c_f v_{f[i]} + c_s v_{s[i]} + c_r v_{r[i]} + c_l v_{l[i]}$$

Weighting factors

Go forward	Stop	Turning right	Turning left
$c_f = -2$	$c_s = 1$	$c_r = 2$	$c_l = 2$

3. Estimation of Importance

- Normalize importance
- Regard crosswalk section
  - Total output is over pre-defined threshold
- Add bias to crosswalk section

$$S''_i = \min(S'_i + 0.5, 1.0)$$

4. Calculation of Playing Speed

- Calculation play speed from importance

$$sp[i] = \frac{1}{S''_i\left(1 - (1/(sp_{max}))\right) + (1/(sp_{max} - 1))} + 1$$

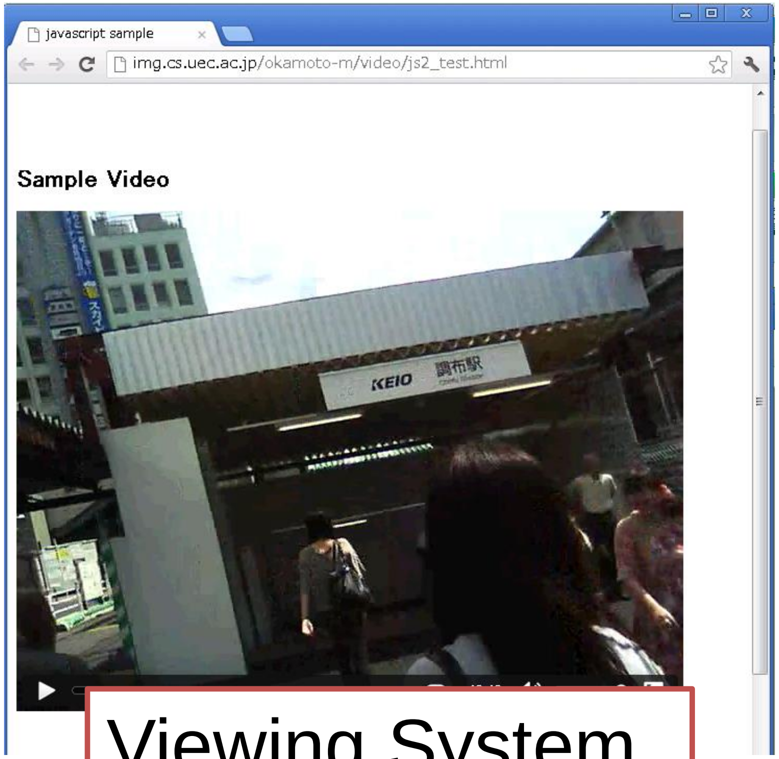
- sp_{max} is given by user when playing
  - User can adjust max playing speed on-line

- Smoothing playing speed (for easy watching)

$$sp'[i] = 0.1(sp[i-1] + sp[i+1]) + 0.8\, sp[i]$$

Viewing System

- Implemented view system in HTML5
  - Be embed classifier outputs and numbers of detection of each video section

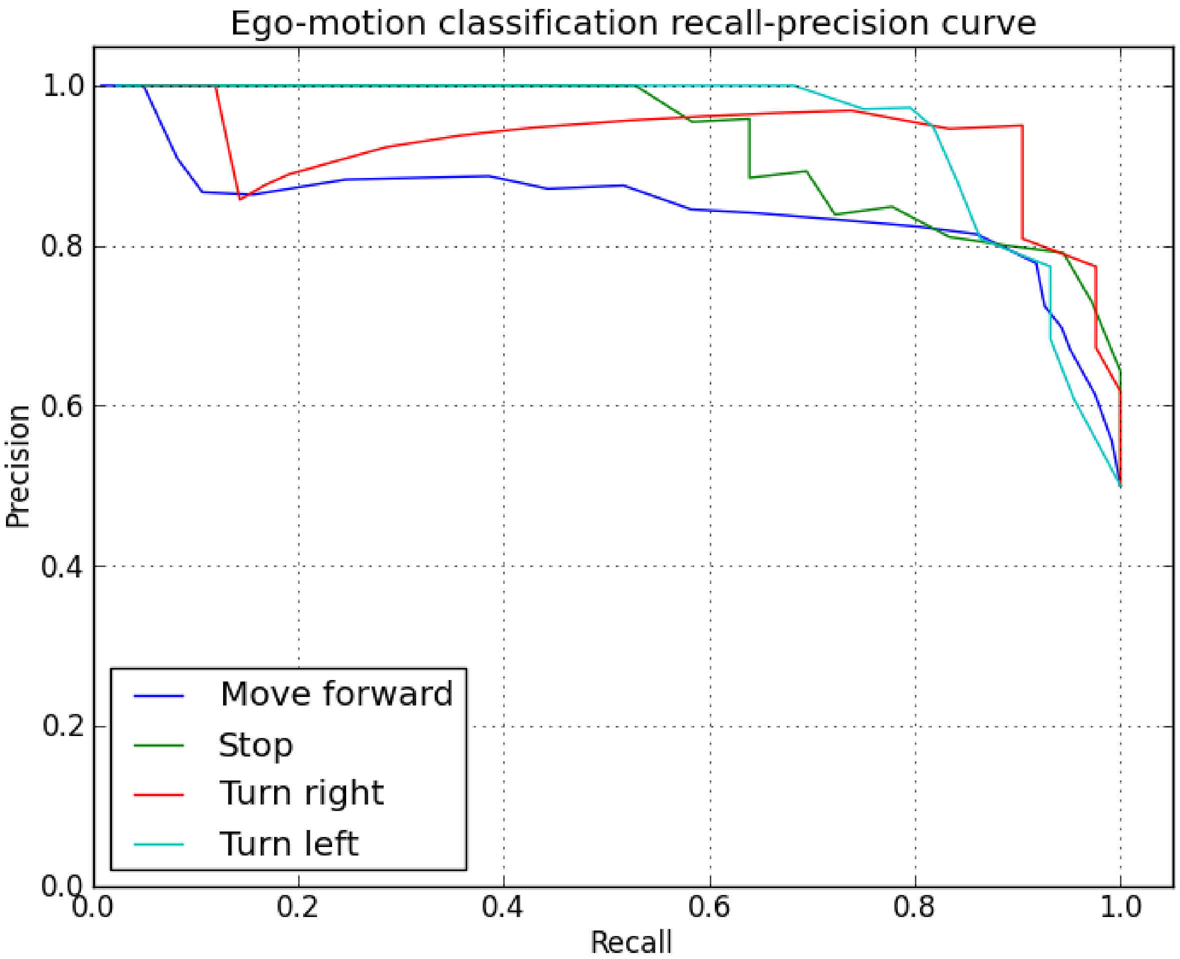


Viewing System

Experiments

- Dataset
  - – Taken at around our university (Tokyo)
  - – 9 Videos (average 9min long)
- Evaluation experiments
  - – Ego-motion classification
  - – Crosswalk Detection
- User study
  - – Vote best summarization method by users

Evaluation of Ego-motion Classification



Classification rate ((TP + TF) / N) = 83.8 %

Evaluation of Crosswalk Detection

- Experiment Setup
  - – 250 learning frames from four videos
  - – 200 testing frames from five videos
- Compare w/ and w/o ground estimation
  - – To evaluate the ground region estimation

Evaluation of Crosswalk Detection

	Recall	Precision	f-number
w/ ground estimation	0.37	0.787	0.503
w/o ground estimation	0.26	0.839	0.397

Improvement of F-number 0.106

User Study Setup

Comparing methods as follows:

① Proposed (Ego-motion + Crosswalk)

② Ego-motion classification only

③ Fast-forwarding at a uniform speed

④ Storyboard-style

Ask 10 subjects to evaluate the results for three videos by above

User Study Result

- Most of the subjects voted to proposed method

Video	Ego-motion	Ego. + crosswalk	Fast-forwarding	Storyboard
Video A	4	6	0	0
Video B	3	6	1	0
Video C	1	7	1	1
Total	8	19	2	1

Conclusion

- Summarization of egocentric moving videos for generating route guide videos
- Experiment result
  - Achieve 83.8% in ego-motion classification
  - Ground estimation improve crosswalk detection
- User study
  - Most subject voted proposed method

Future Works

- Extending target videos
  - Focused on only walking videos now
  - Bike and car egocentric videos
- Adding important objects
  - Use other object cues for deferent situation

Contacts

- Masaya Okamoto
- e-mail: okamoto-m@mm.inf.uec.ac.jp

Gracias!

VS Google Street Views

- Taken from high view point
- Limited to large cities

Ours

Street Views

Expression of Playing Speed

- Calculation play speed from importance

$$sp[i] = \frac{1}{S''_i \left(1 - (1/(sp_{max}))\right) + (1/(sp_{max} - 1))} + 1$$

- s_{max} is given by user when playing
  - User can adjust max playing speed on-line

Target Video

We assume our target videos are

① Walking video recorded from a starting place to a destination

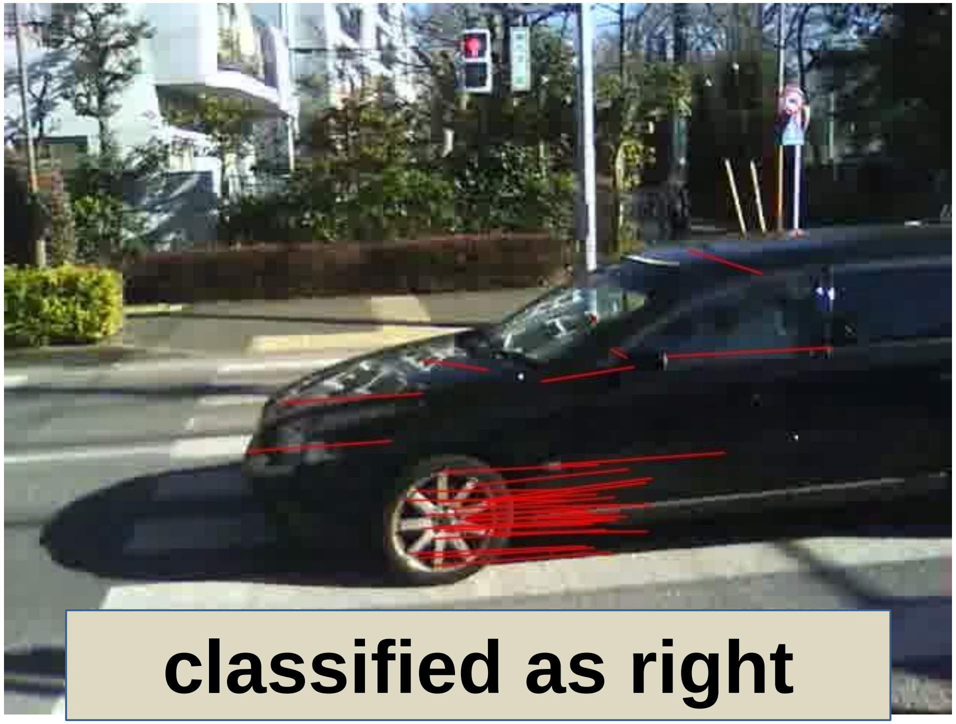
② Recorded by a moving wearable camera

③ Recorded continuously (not interrupted)

A
調布市
石原小
文
北多摩病院
市
12
20
B
調布
調布市役所

Optical Flows from Moving Object

- Optical flows from moving object (ex. Car) cause failure of ego-motion classification



classified as right

Calculation of Optical

- We use improved LK module in OpenCV
- For high accuracy

Detail Setup of User study

- Use 3 videos for User study
- Taken by me at residential area Tokyo

Video	Duration	After duration	Average speed	Storyboard size
Video A	7:47	1:45	4.5	21
Video B	9:17	2:20	3.9	28
Video C	11:26	2:40	4.3	32

Learning of weight factors

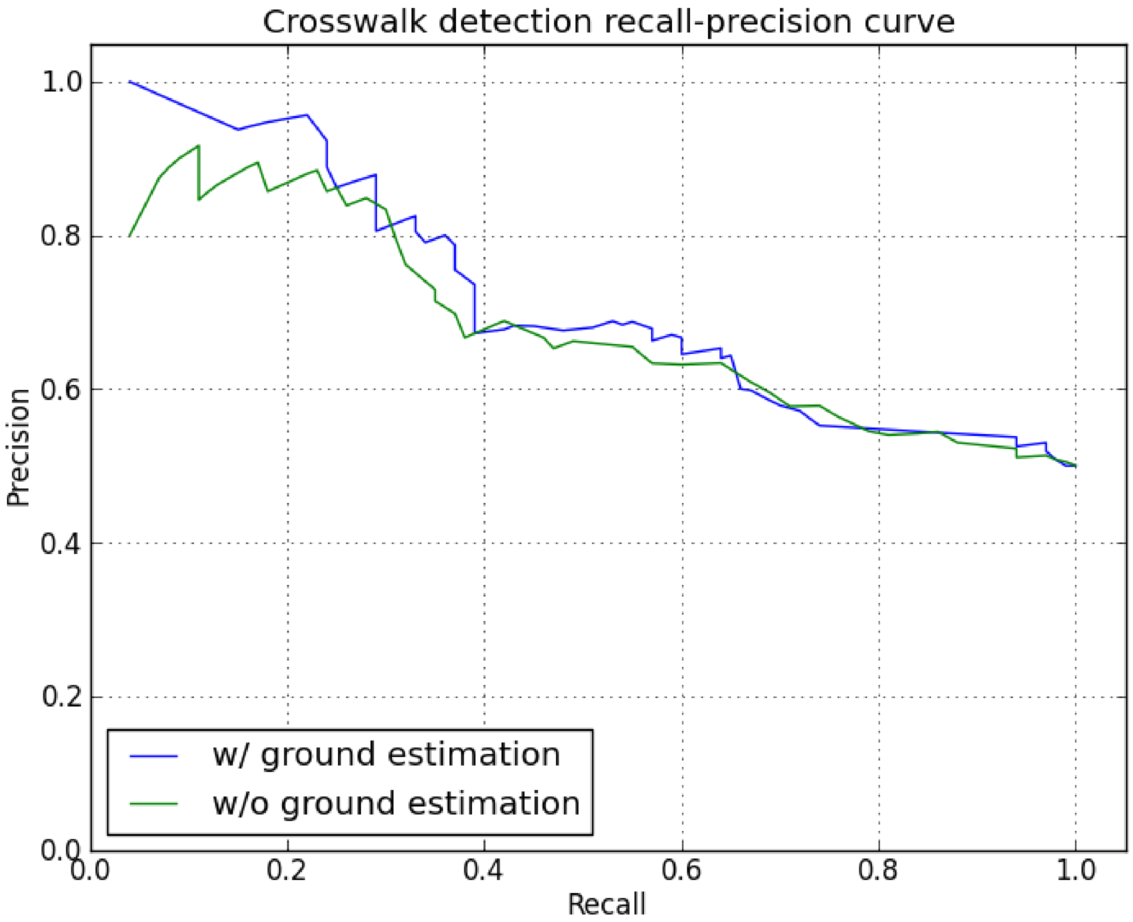
- Decided by preliminary experiments
- We will decide several parameters from training data
  - Need supervised signals in each video section

Why not a multi-class classifier

- Some section contains complex motion
  - Two motions in one section, looking aside
- One-vs-all classifiers can represent complex motion

Evaluation of Crosswalk Detection

Recall-precision curves



Recall-Precision Curve of Crosswalk detection

2. Crosswalk Detection

- Extract SIFT feature from ground regions
- Make BoF vector with extracted SIFT
- Use non-linear SVM as classifier
  - about 240 learning frames

Our dataset

- Collected videos contain somebody faces and car numbers

- It’s difficult to distribute