Daily Activity Recognition Combining Gaze Motion and Visual Features

Yuki Shiga, Takumi Toyama, Yuzuko Utsumi,
Andreas Dengel, Koichi Kise

大阪府立大学
OSAKA PREFECTURE UNIVERSITY

Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

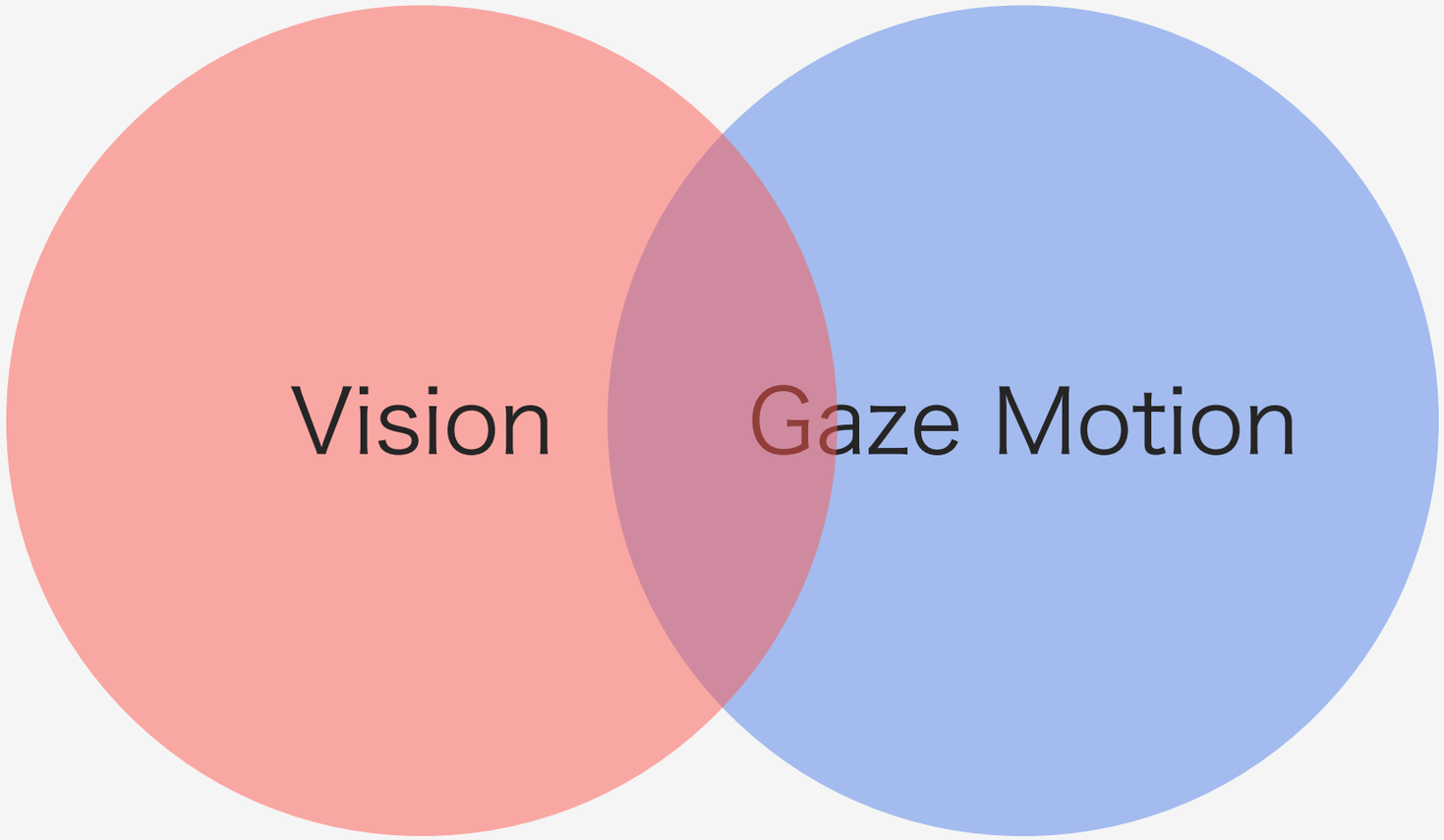
Outline

- Introduction
- Proposed Method
- Experiment
- Conclusion

Outline

- **Introduction**
- Proposed Method
- Experiment
- Conclusion

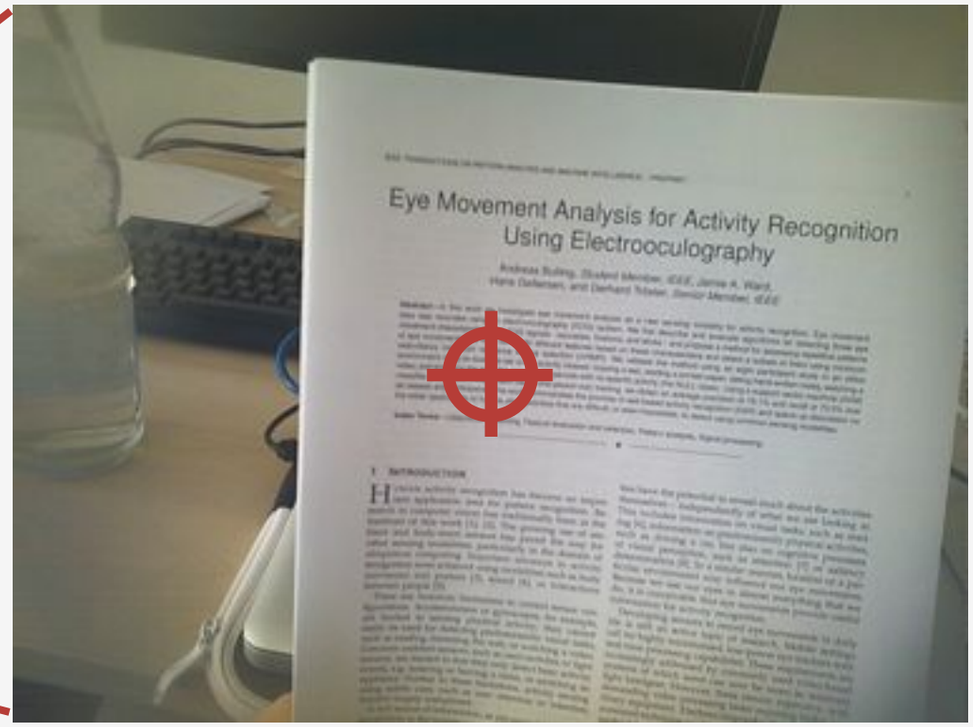
Focus



- Activity recognition draws public attention
- Focus on vision-based and Gaze motion-based method
- These methods deal with activities that involve eye movements

Eye Tracker

Gaze Position
(Where the User Fixates)



Scene Image

- An eye tracker is useful for recognizing activities that involve eye movements
- Record a scene image video as well as the gaze position data

Related Works

- Gaze motion-based activity recognition:
  - Bulling et al., “Eye movement analysis for activity recognition using electrooculography.”[1]
- Vision-based activity recognition:
  - Hipny et al., “Recognizing Egocentric Activities from Gaze Regions with Multiple-Voting Bag of Words.”[2]

They used only each modality (Motion or Vision)

[1] Bulling, Andreas, Ward, Jamie, Gellersen, Hans, and Töster, Gerhard. Eye movement analysis for activity recognition using electrooculography. IEEE transactions on pattern analysis and machine intelligence, 33, 4 (2011), 741-53.

[2] Hipiny IM, Mayol-Cuevas W. Recognising Egocentric Activities from Gaze Regions with Multiple-Voting Bag of Words. CSTR-12-003. 2012.

Purpose

Activity

- can be expressed by "how eyes move”
- can also be expressed by "what eyes see”

We use both vision-based and gaze motion-based modality for activity recognition

Purpose

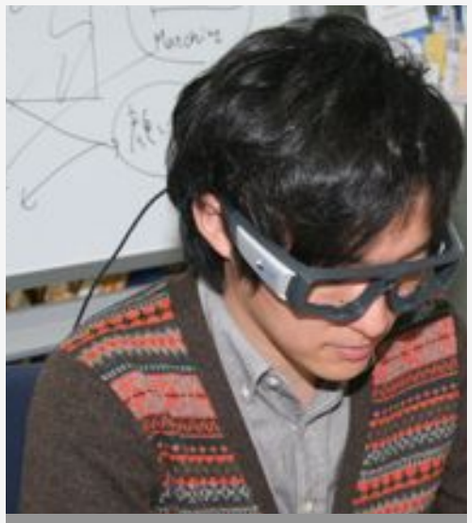
- Propose a method combining gaze motion-based method and vision-based method

- Verify the hypothesis:
  Both combination of vision and gaze motion can improve recognizing activities that involve eye movements

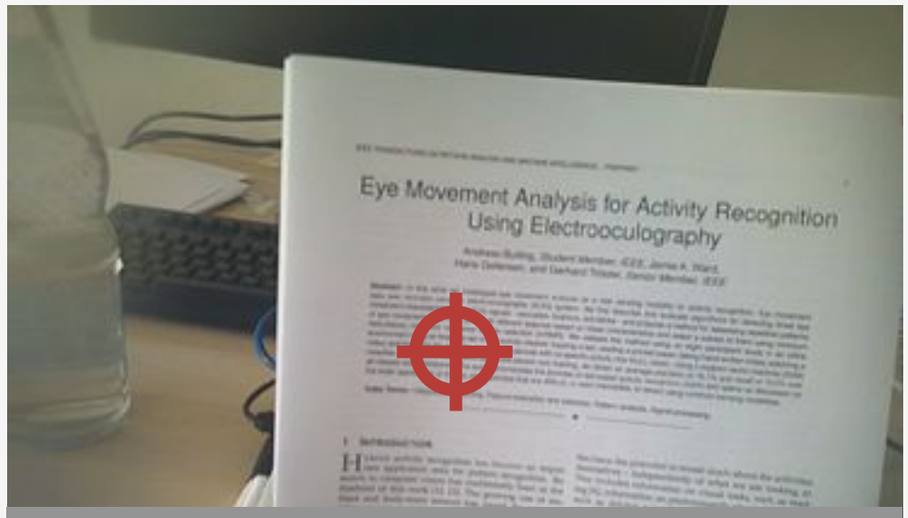
Outline

- Introduction
- Proposed Method
- Experiment
- Conclusion

Overview



Eye Tracker



Record Gaze Points
and Scene Images

Gaze Motion
Feature

Visual Feature

Classifier

Classifier

Output

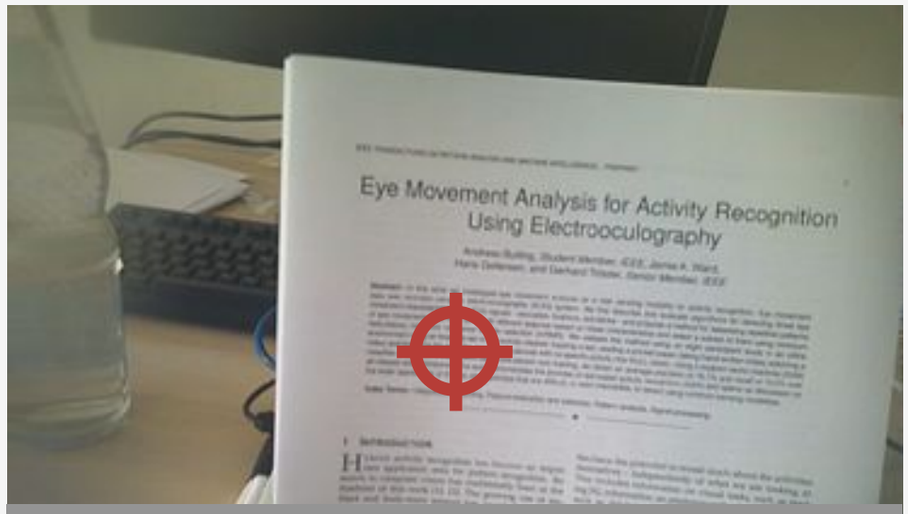
Output

Fusion

Result

Overview

Eye Tracker



Record Gaze Points
and Scene Images

Gaze Motion
Feature

Visual Feature

Classifier

Classifier

Output

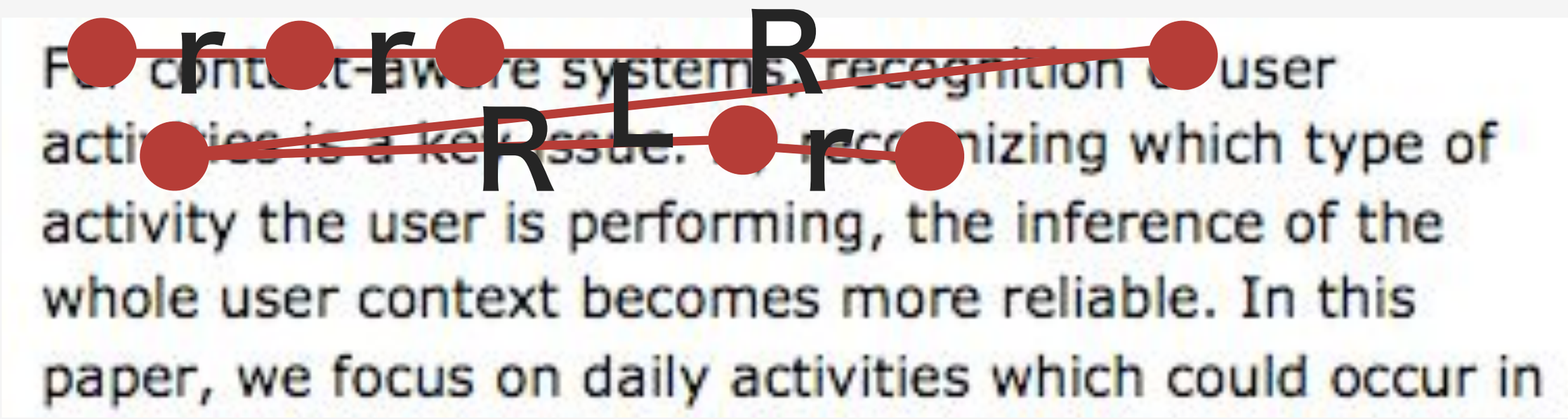
Output

Fusion

Result

Gaze Motion Feature

- The method proposed by Bulling et al.



Saccade

Fixation

Convert

rrRLRr

Representing Size and Direction of Saccade

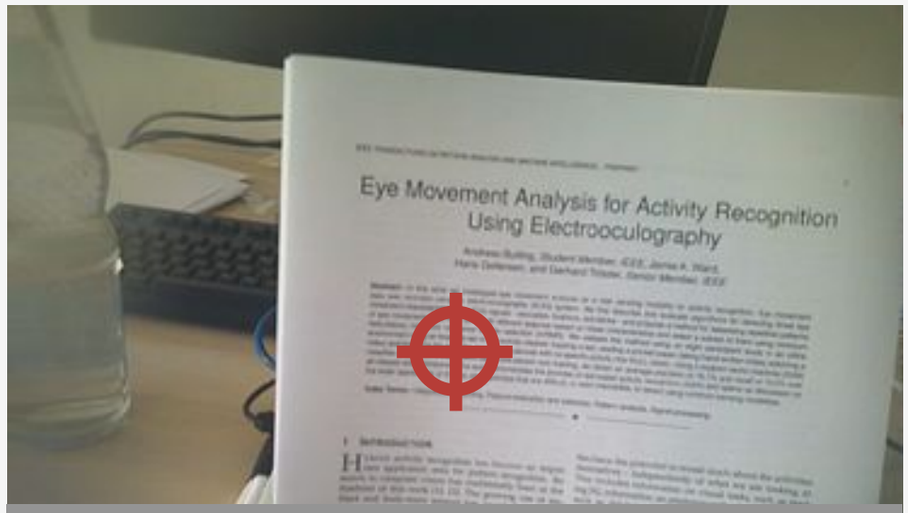
N-gram method

Statistical Fature

[1] Bulling, Andreas, Ward, Jamie, Gellersen, Hans, and Töster, Gerhard. Eye movement analysis for activity recognition using electrooculography. IEEE transactions on pattern analysis and machine intelligence, 33, 4 (2011), 741-53.

Overview

Eye Tracker



Record Gaze Points
and Scene Images

Gaze Motion
Feature

Visual Feature

Classifier

Classifier

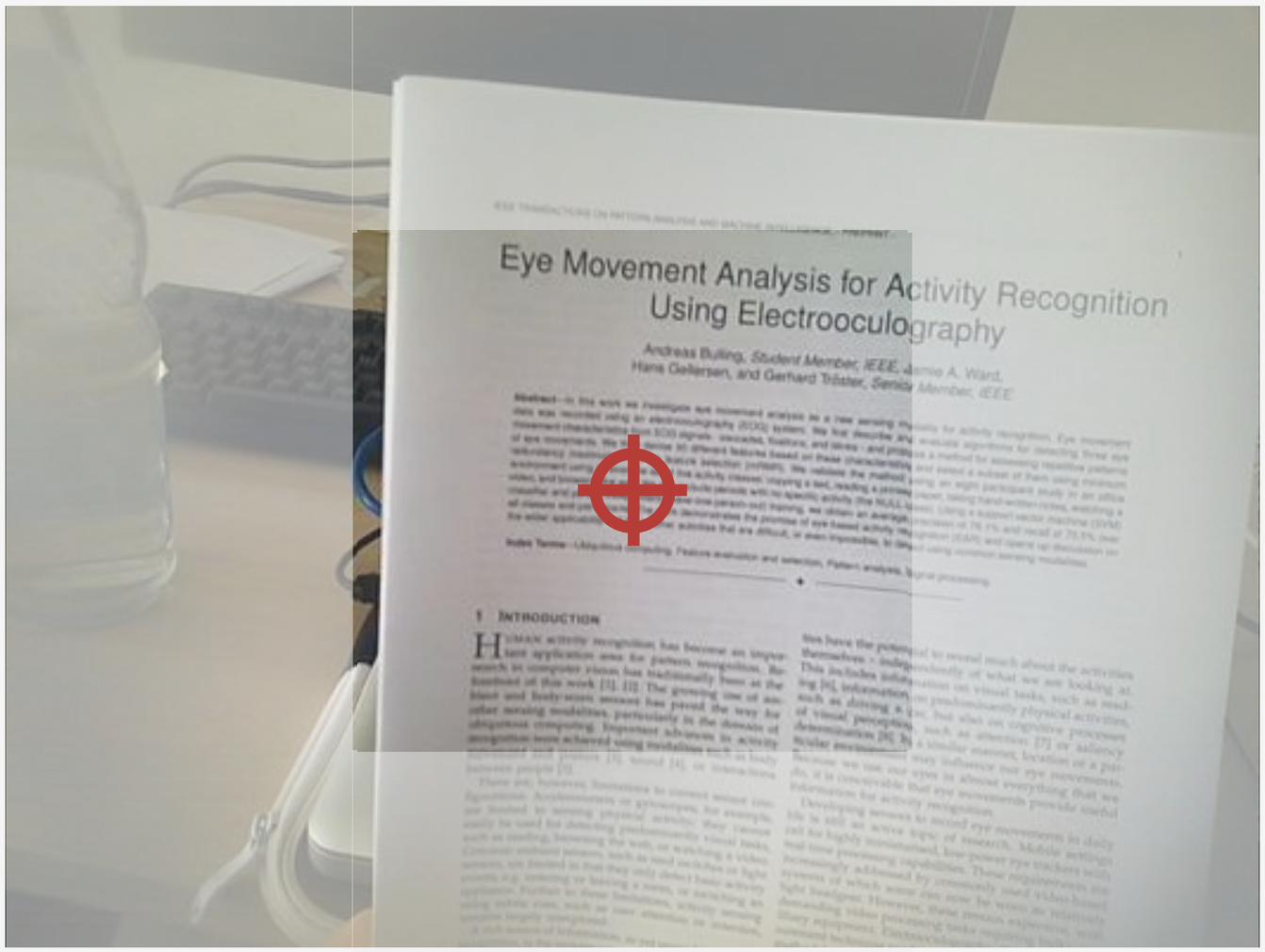
Output

Output

Fusion

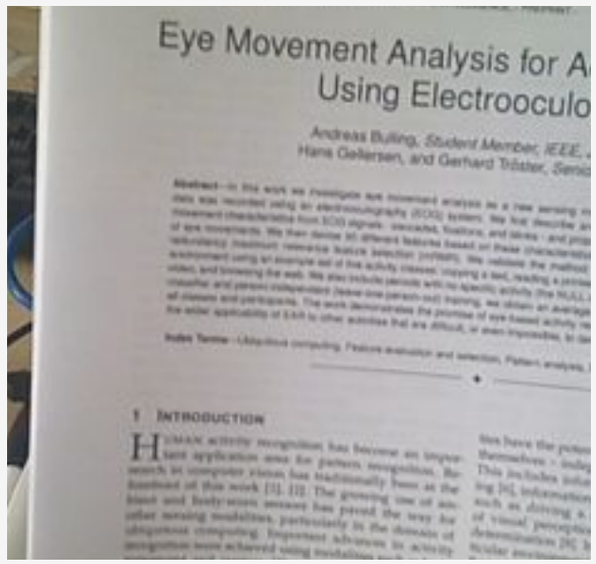
Result

Visual Feature



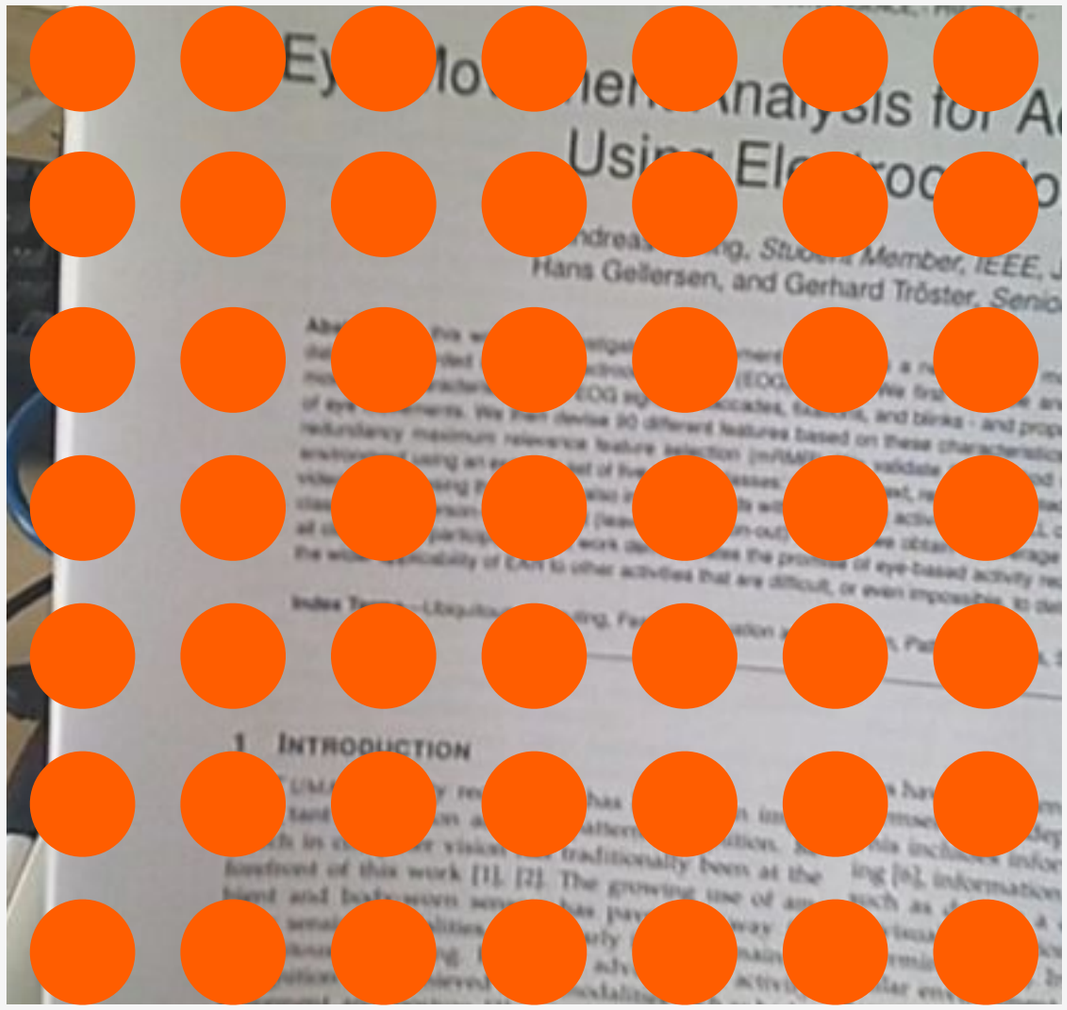
Crop a region around gaze points
to remove a irrelevant region

Visual Feature



Crop a region around gaze points
to remove a irrelevant region

Local Feature Extraction

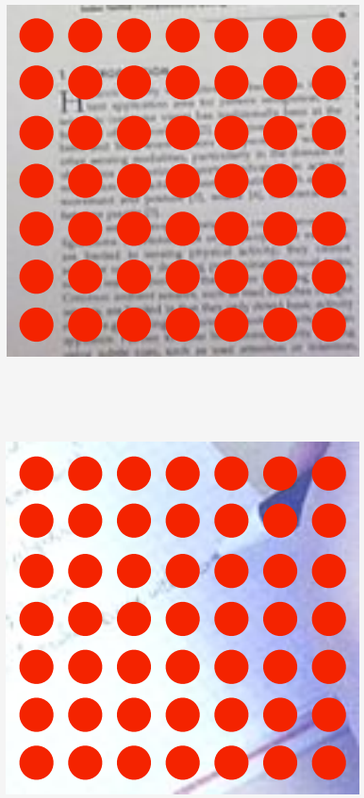


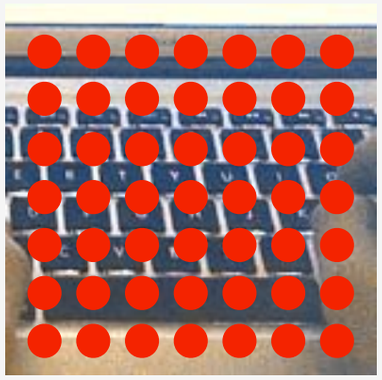
Intrest Points
by Dense Sampling

Extract Local Features
(PCA-SIFT)
From Each Point

Convert to Global Feature

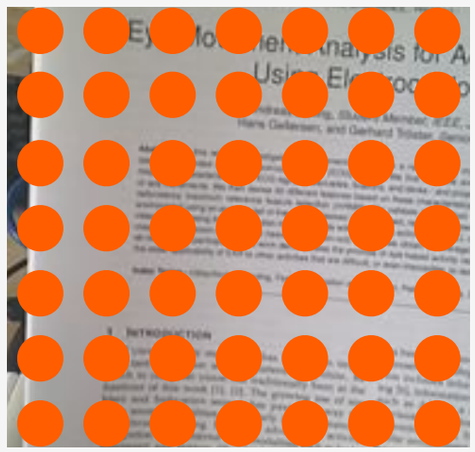
Learning Image





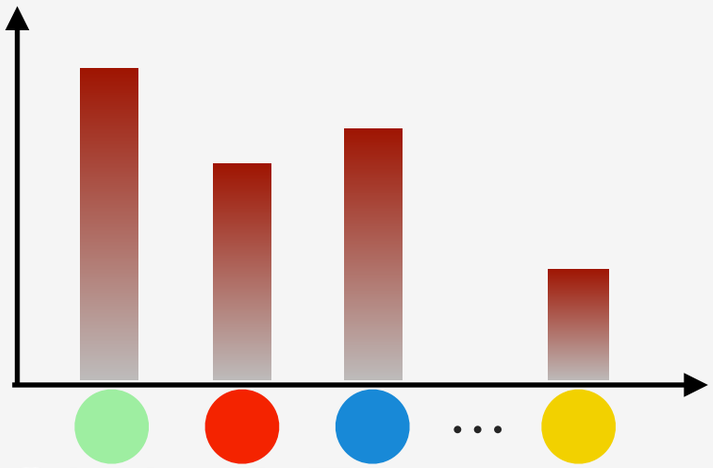
k-means clustering

k centroids (visual words)



Test Image

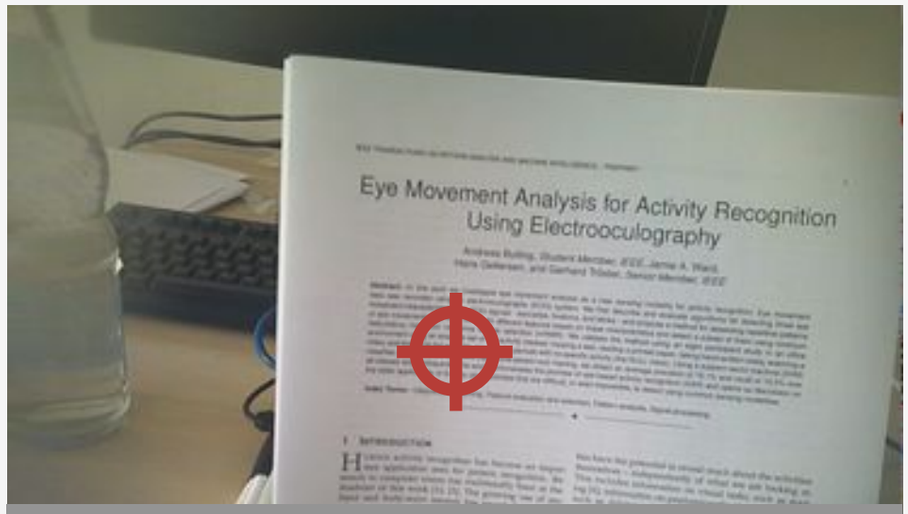
Nearest Neighbor Search to visual words



Global Feature

Overview

Eye Tracker



Record Gaze Points
and Scene Images

Gaze Motion
Feature

Visual Feature

Classifier

Classifier

Output

Output

Fusion

Result

Classifier

Read

Write

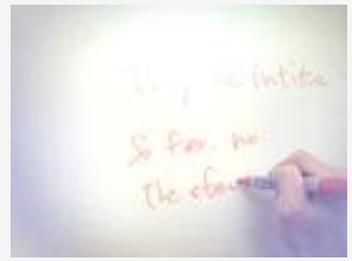
Type

Feature Vector For Learning

- SVM with Probability Estimation
- Two classifiers are made for visual and gaze motion features

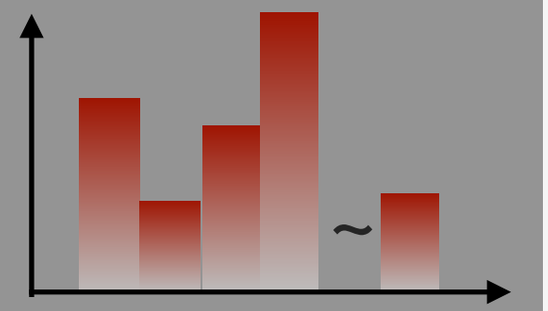
Classifier

Read



Write

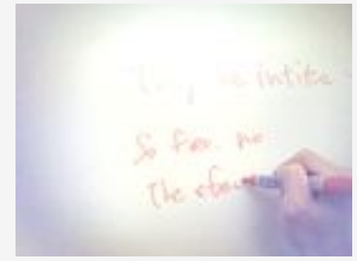
Type



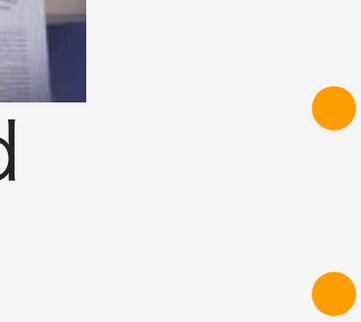
Feature Vector for Test

Classifier

Read



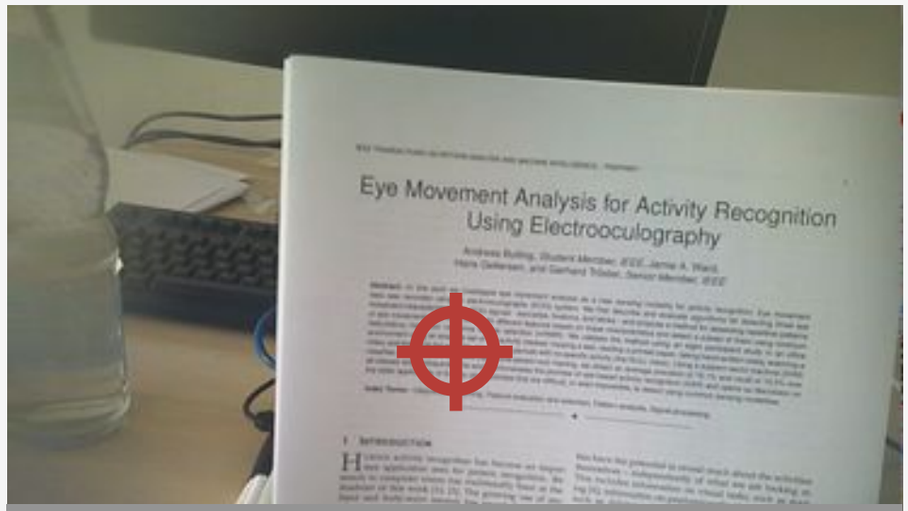
Write



Type

Overview

Eye Tracker



Record Gaze Points
and Scene Images

Gaze Motion
Feature

Visual Feature

Classifier

Classifier

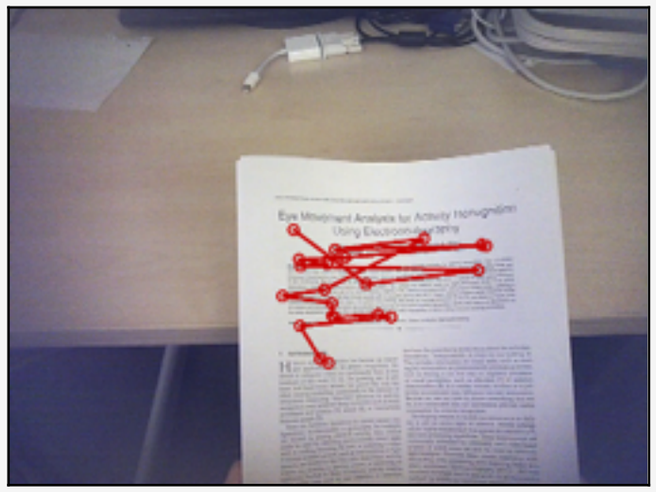
Output

Output

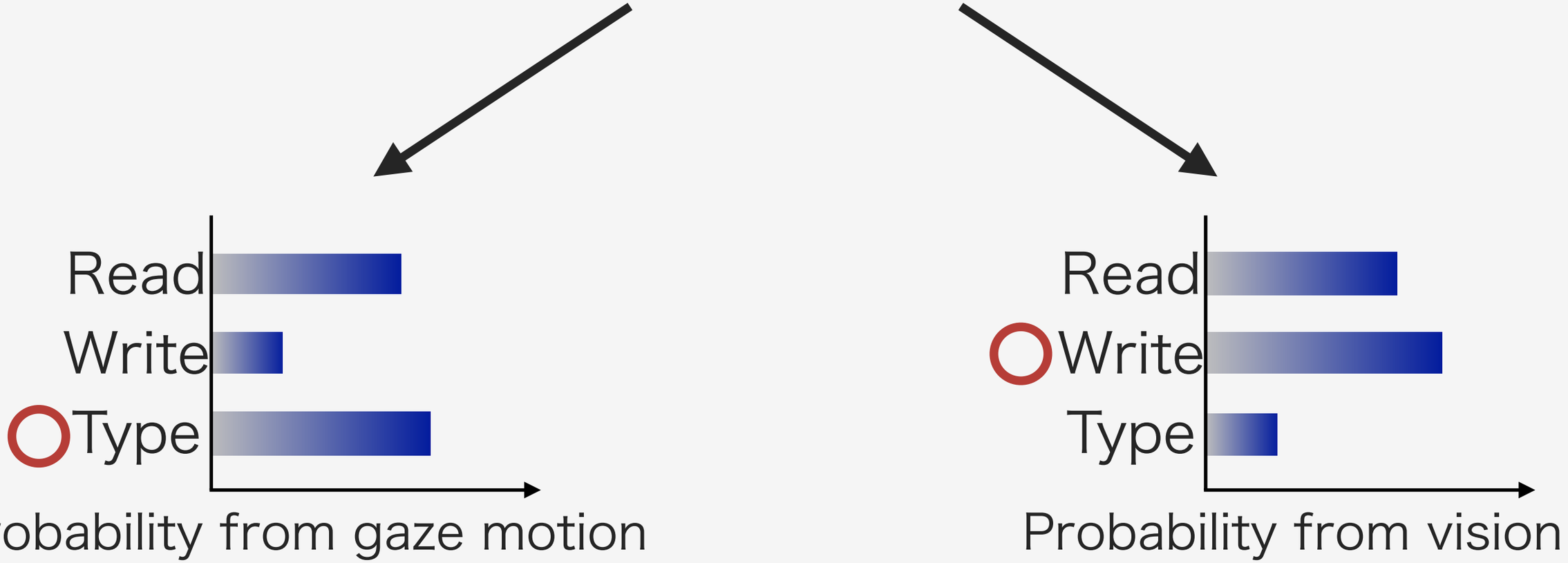
Fusion

Result

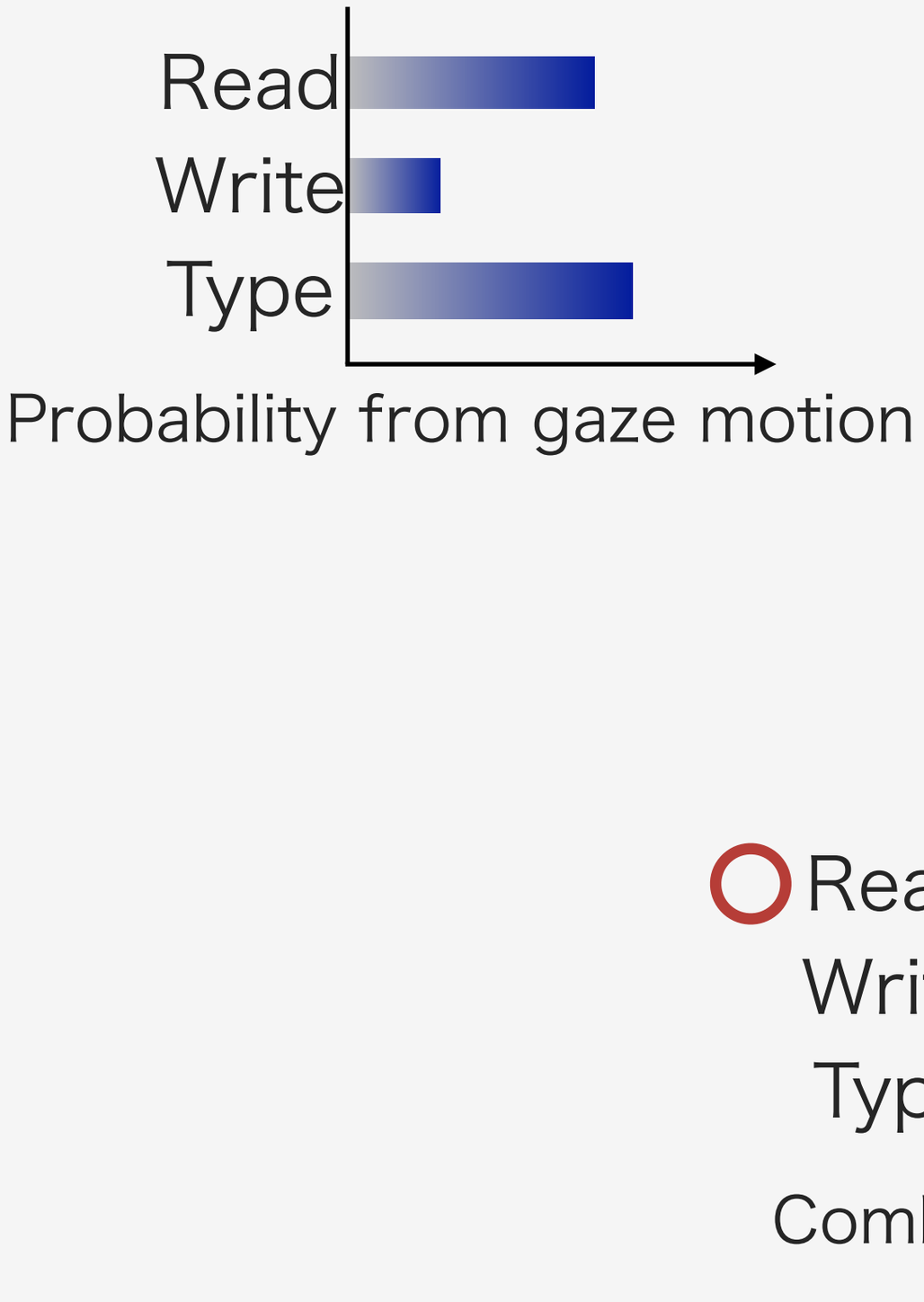
Fusion

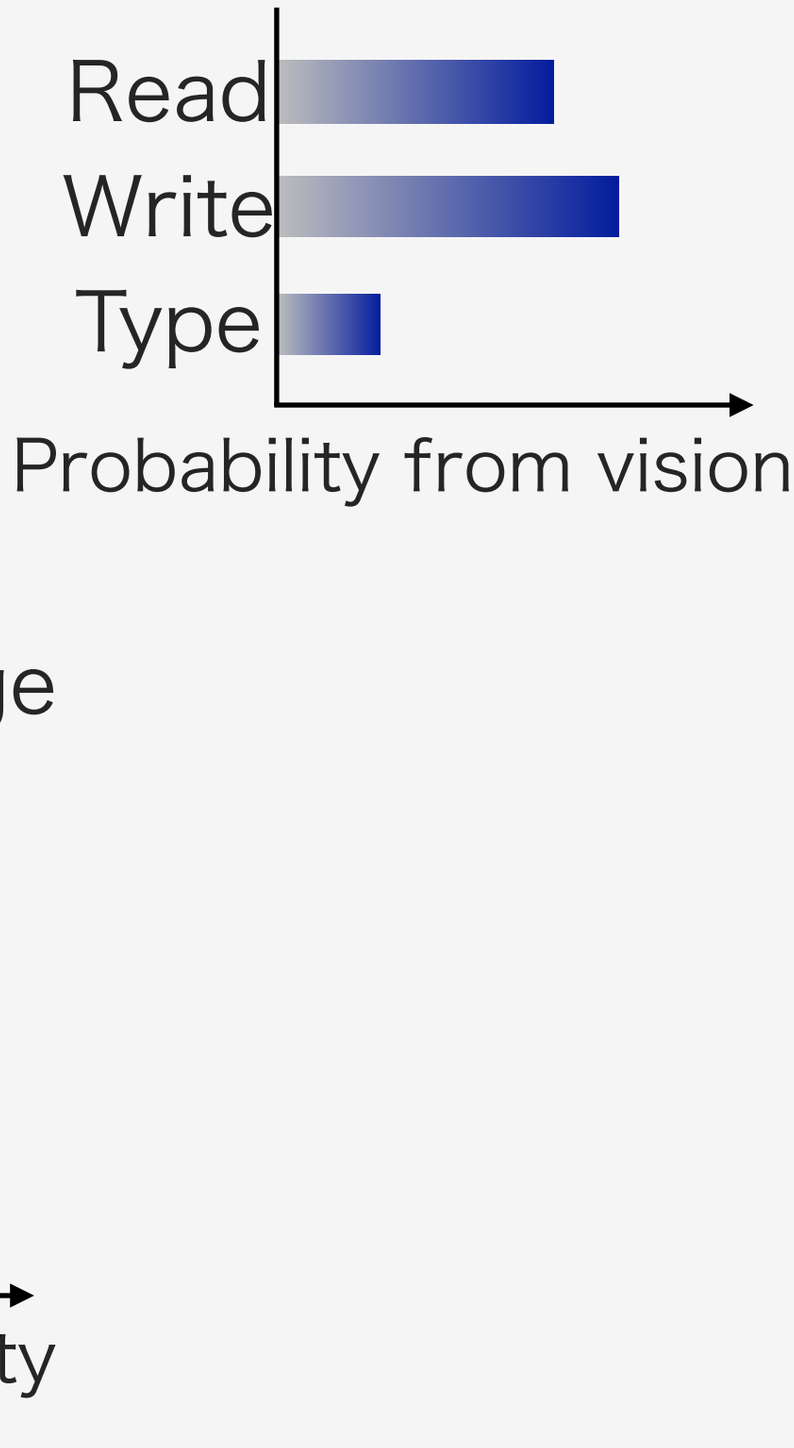


Read

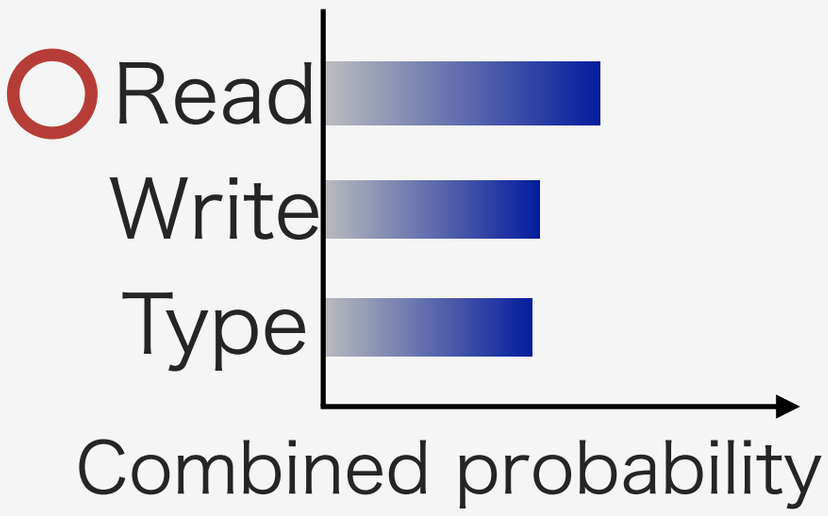

Read
Write
Type
Probability from gaze motion
Read
Write
Type
Probability from vision

Fusion





Average



Outline

- **Introduction**
- Proposed Method
- Experiment
- Conclusion

Experiments

- Baseline:
  Whether combined method performs better than individual vision-based and gaze motion-based method

- Cross-scene:
  Whether the combined method performs when target objects are different between training and test data

- Cross-user:
  Whether the combined method performs when test data contains a person different from training data

	Target Objects / Environments	User
Baseline	Same	Same
Cross-scene	Different	Same
Cross-user	Same	Different

Condition of All Experiments

- Sampling rate of the eye tracker: 30 Hz

- Resolution of the scene camera:
  1280 × 960 Pixels

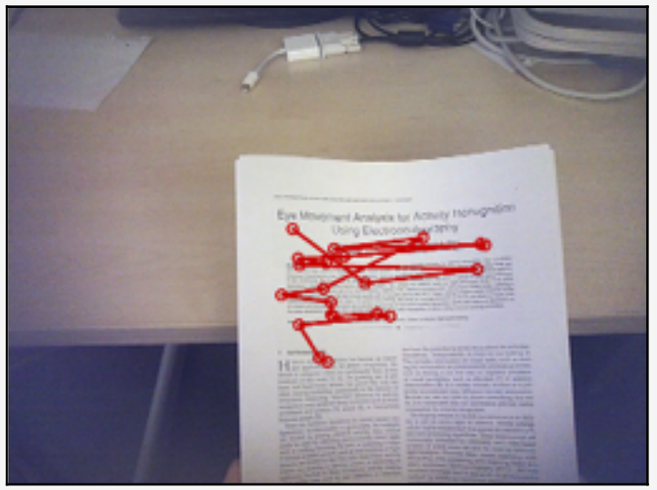
- Visual features are extracted from
  300 × 300 pixels around gaze points

- Gaze motion features are extracted from
  700 gaze samples

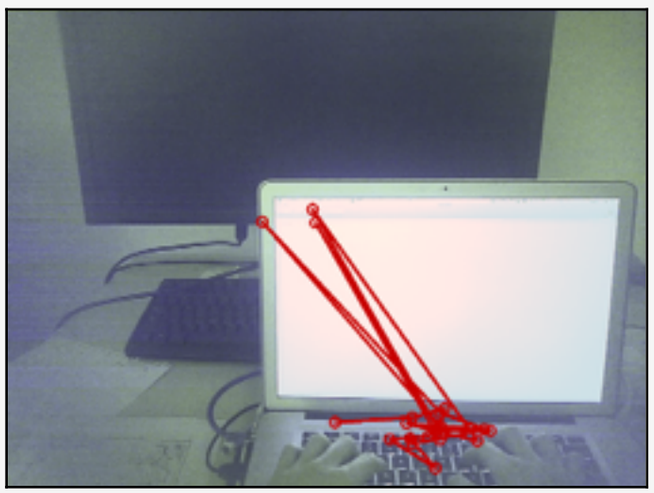
Activity List

Watch a video

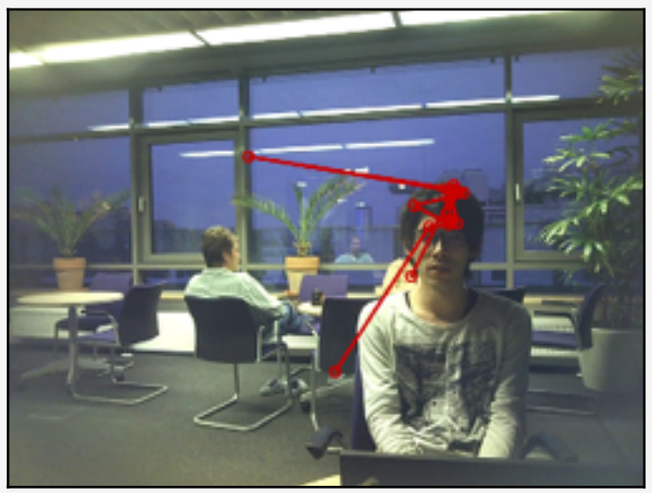
Write text



Read text



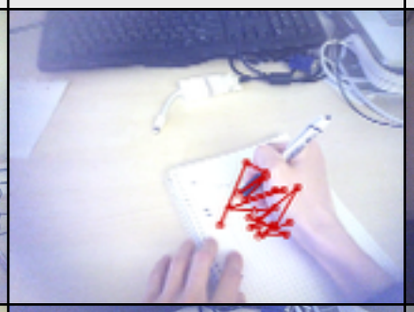
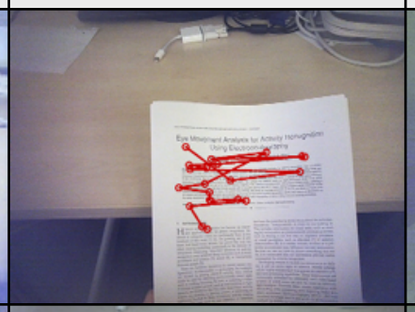
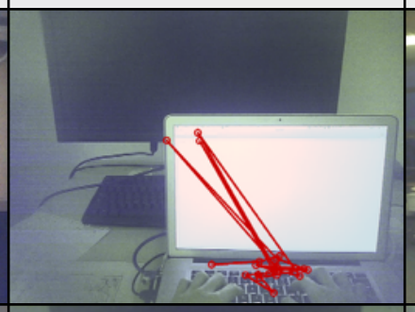
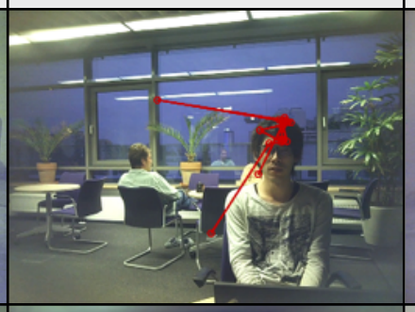
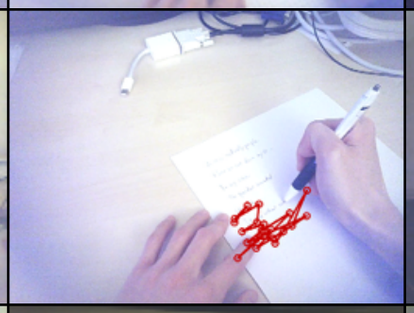
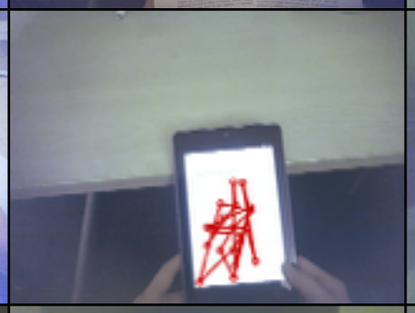
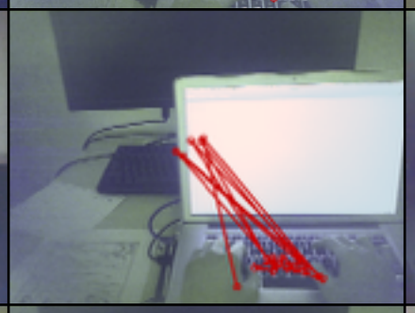
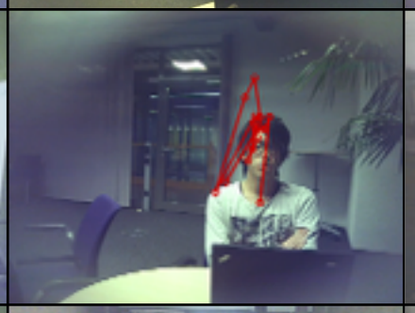
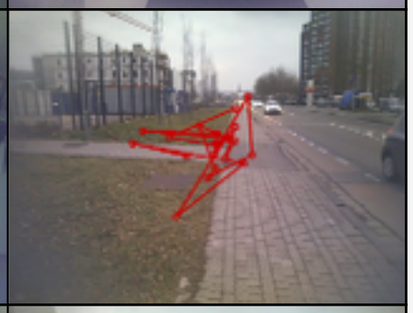
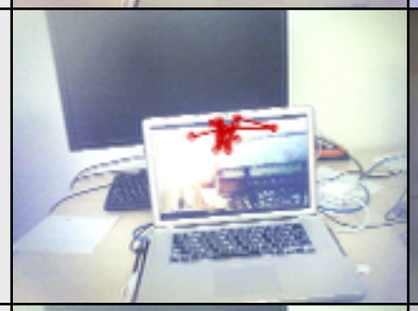
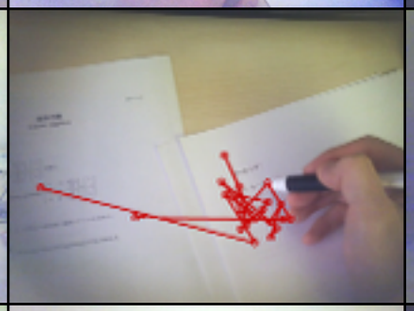
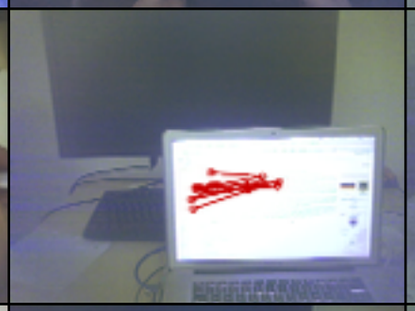
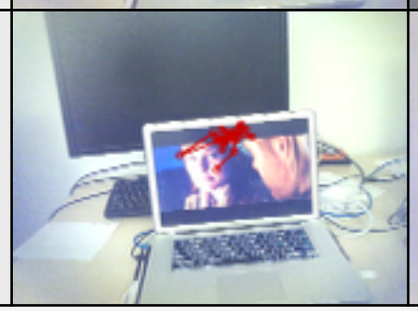
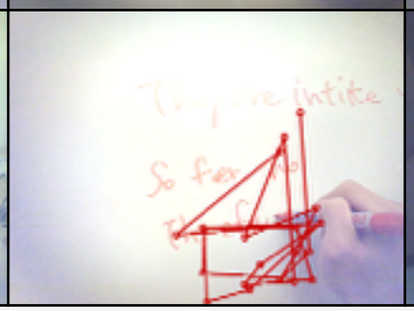
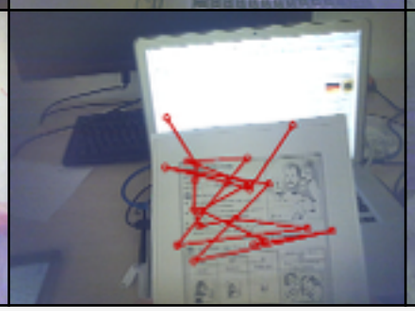
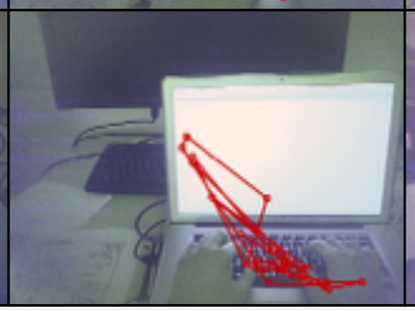
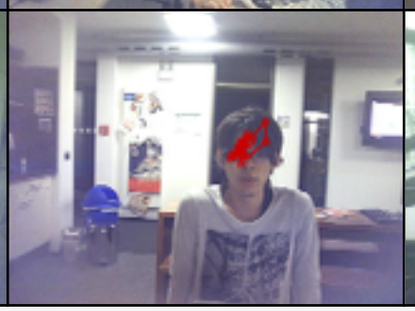
Type text



Have a chat

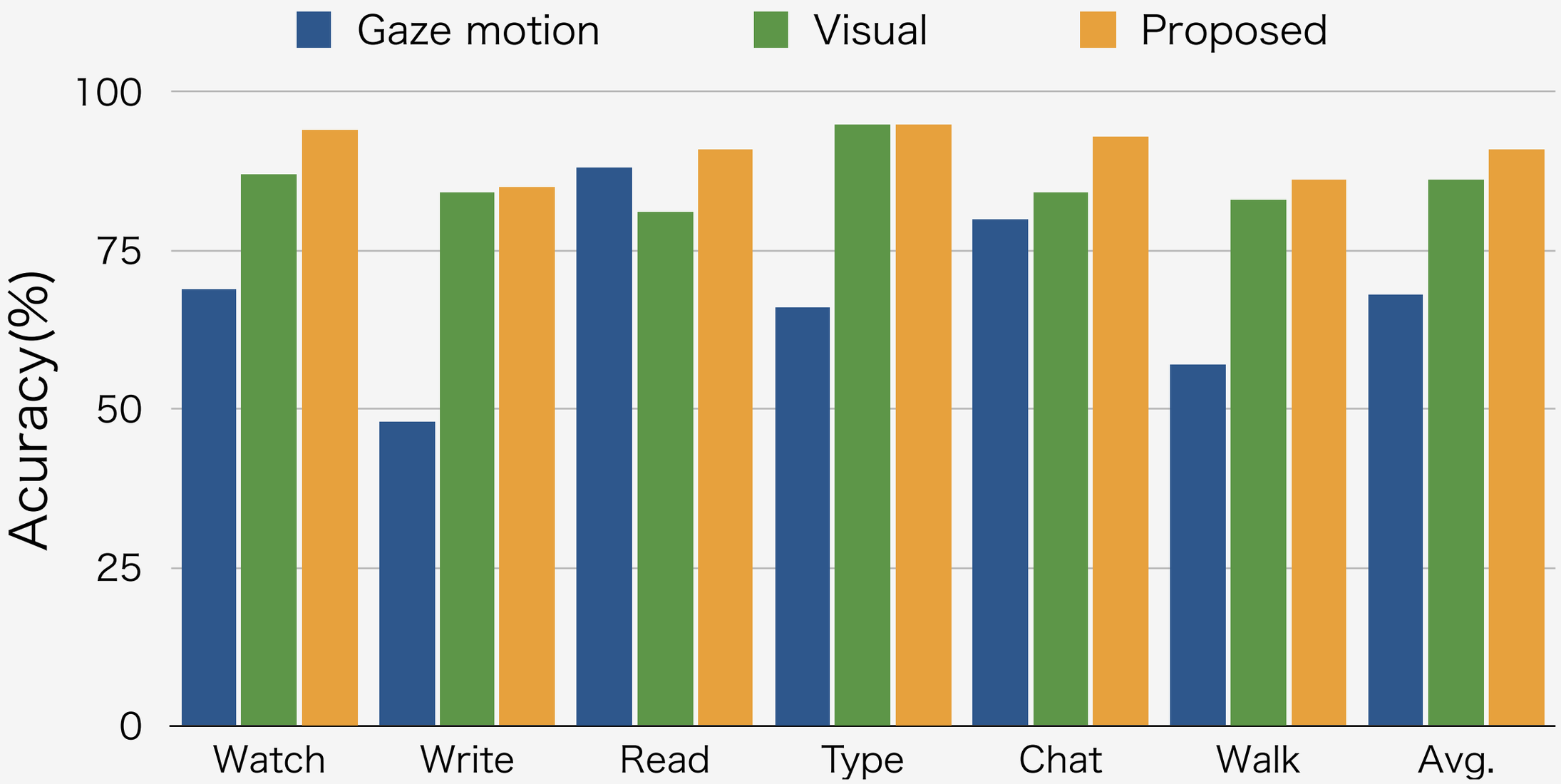
Walk

Baseline Experiment

	Wach a video	Write text	Read Text	Type text	Have a chat	Walk
Scene 1						
Scene 2						
Scene 3						
Scene 4						

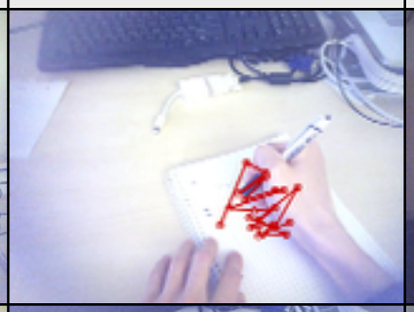
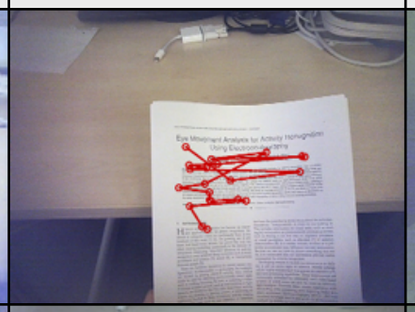
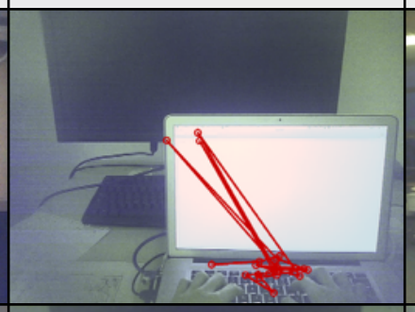
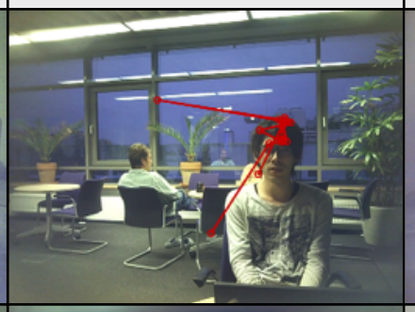
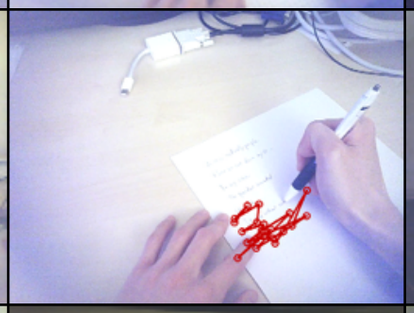
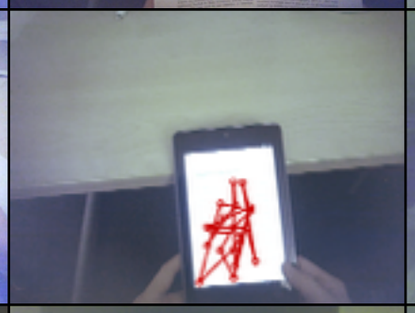
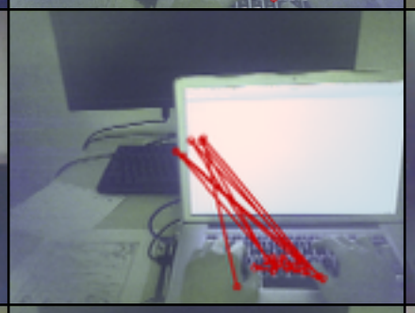
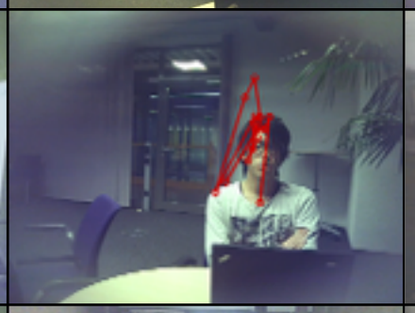
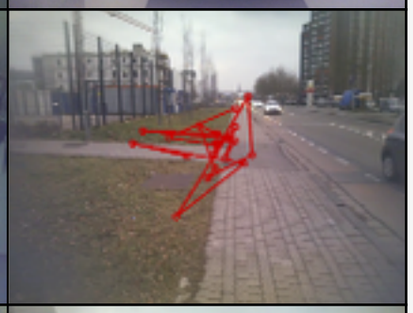
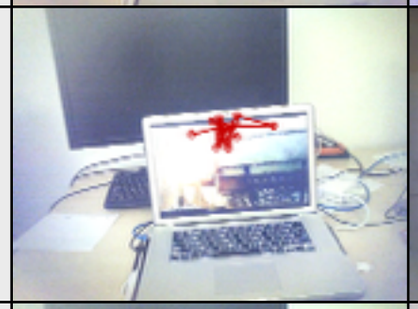
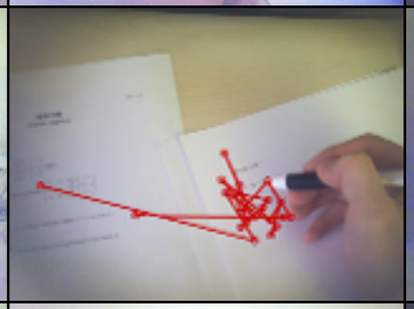
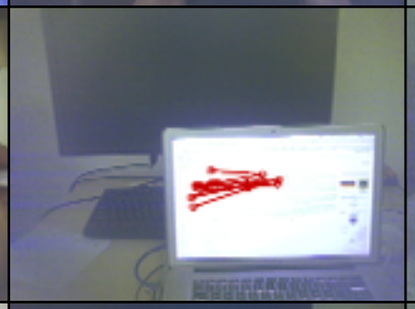
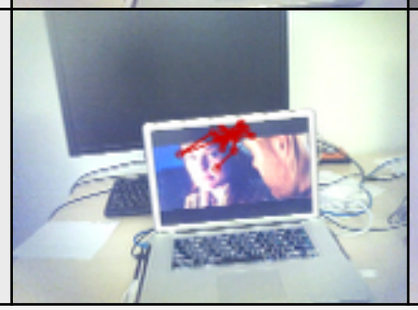
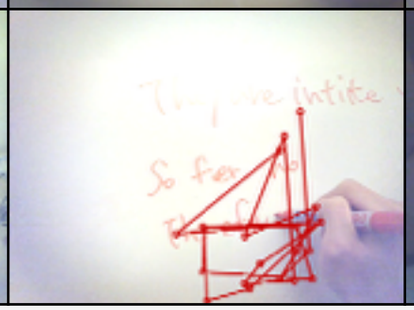
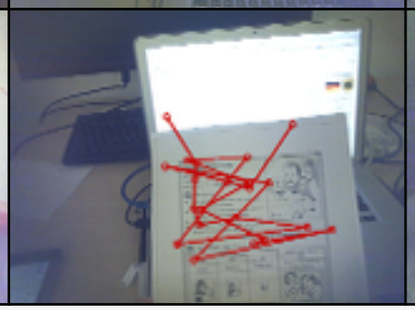
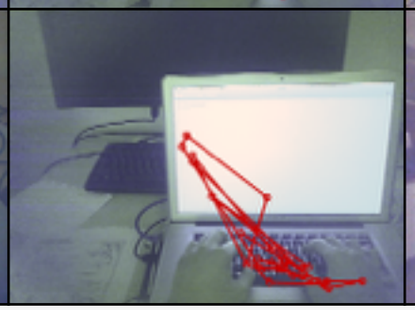
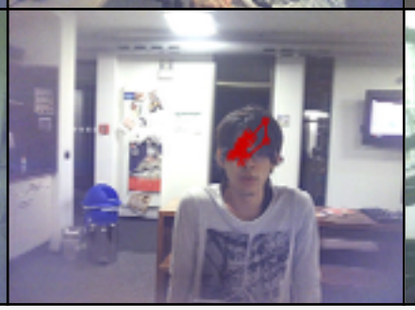
- 1 person
- Contains 4 different scenes
- The dataset was divided into 2 parts

Baseline Experiment



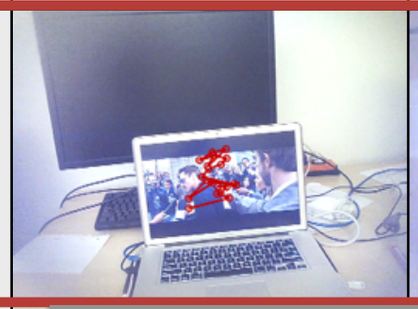
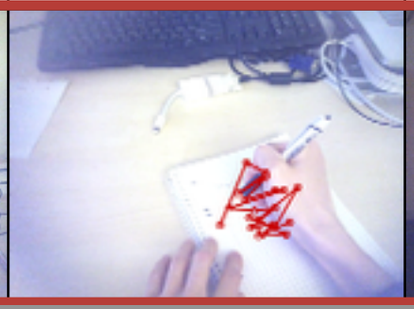
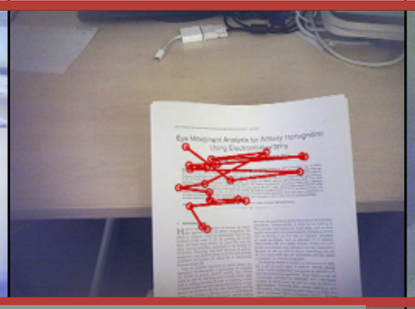
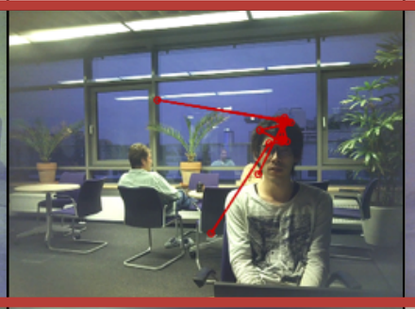
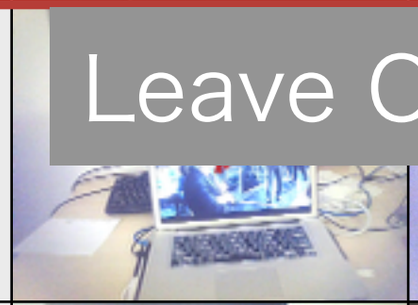
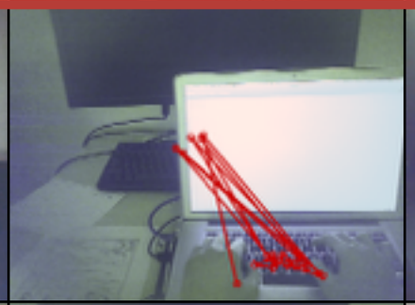
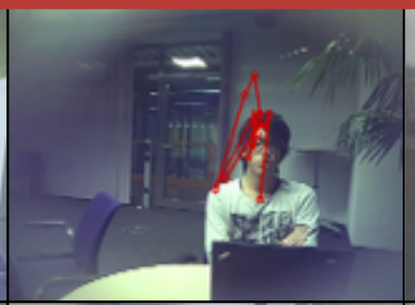
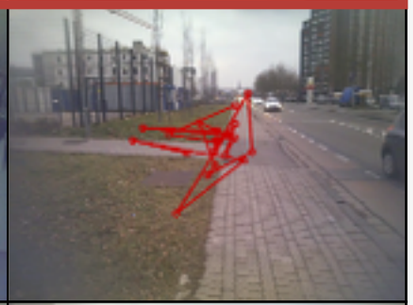
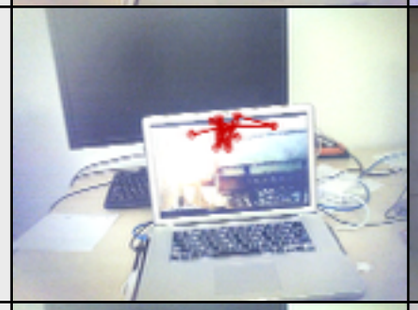
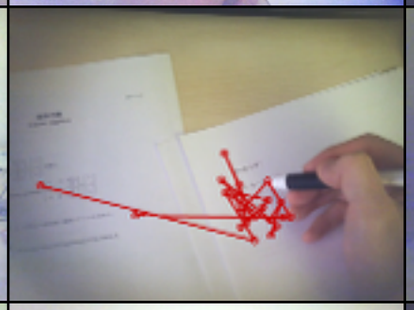
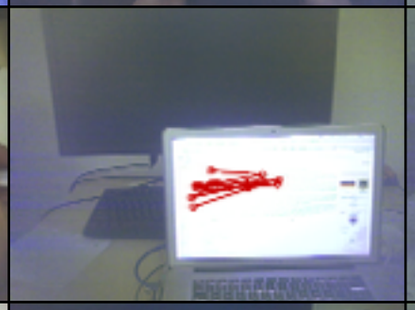
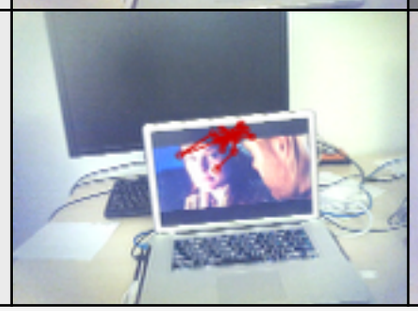
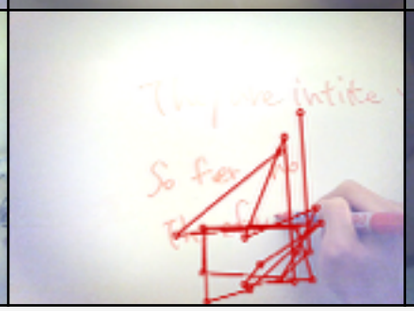
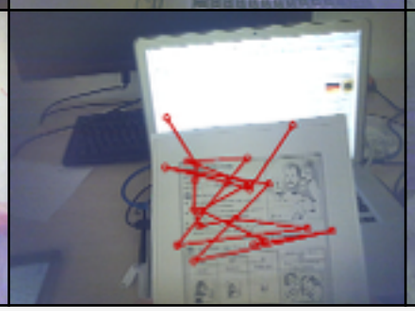
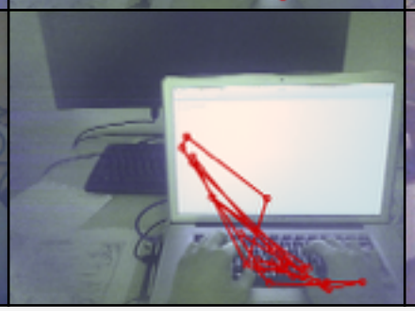
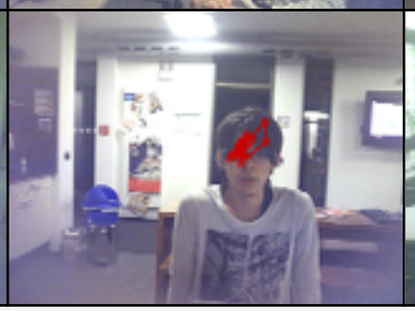
- The accuracy of the proposed method was the best

Cross-scene Experiment

	Wach a video	Write text	Read Text	Type text	Have a chat	Walk
Scene 1						
Scene 2						
Scene 3						
Scene 4						

- 3 people

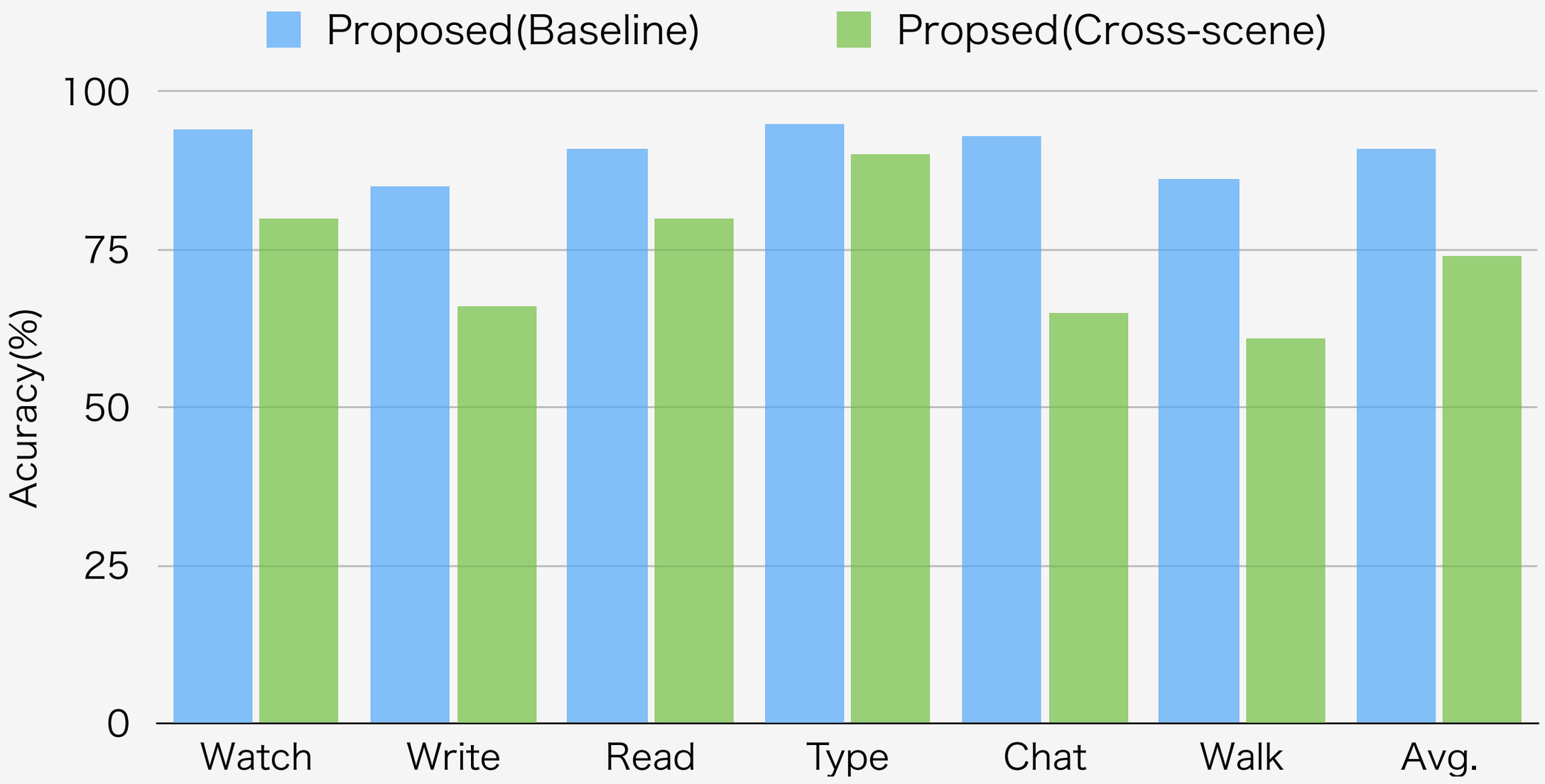
Cross-scene Experiment

	Wach a video	Write text	Read Text	Type text	Have a chat	Walk
Scene 1						
Scene 2						
Scene 3						
Scene 4						

Leave Out for Test Data

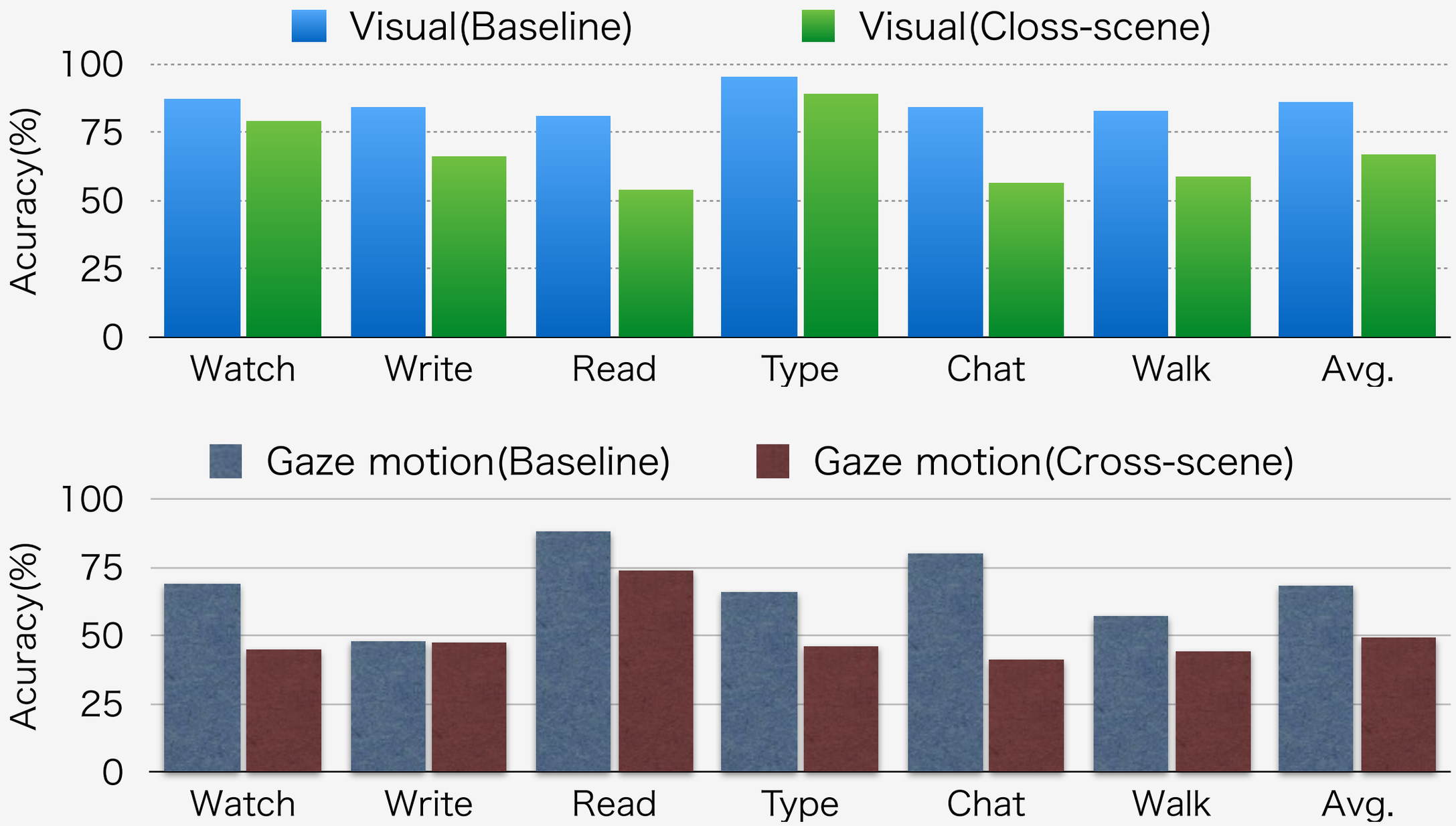
- 3 people
- Leave-one-out cross validation

Cross-scene Experiment



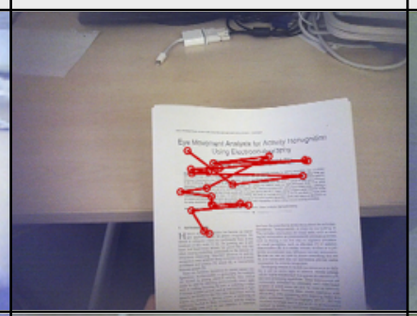
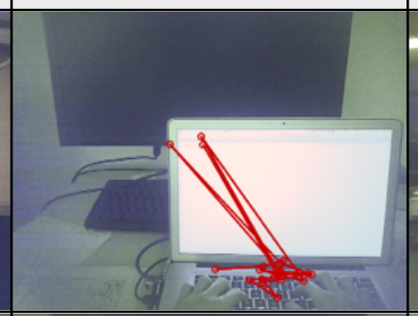
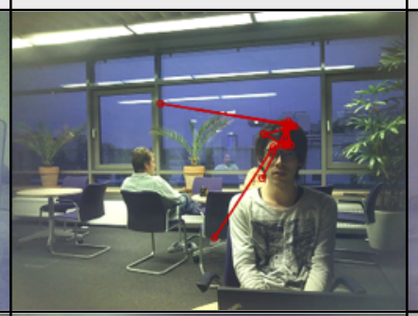
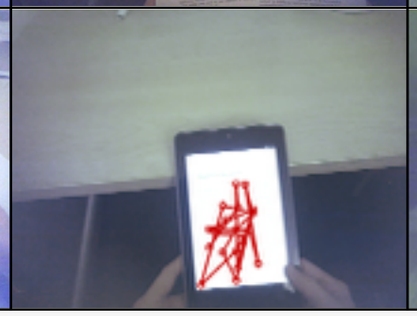
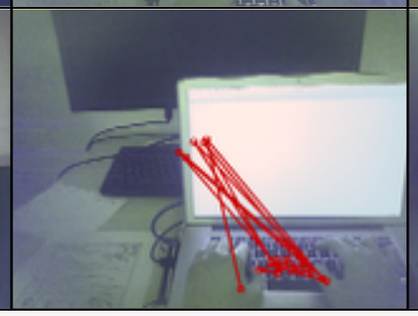
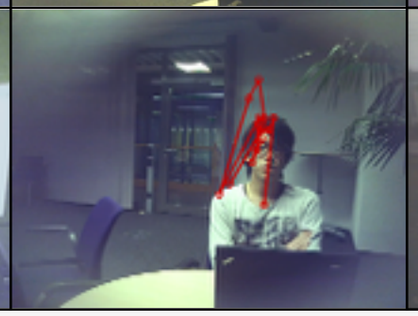
- The recognition rate of Cross-scene is lower than Baseline

Cross-scene Experiment



- Both of recognition rates dropped
- Gaze motion also depends on targets or environments

Cross-user Experiment

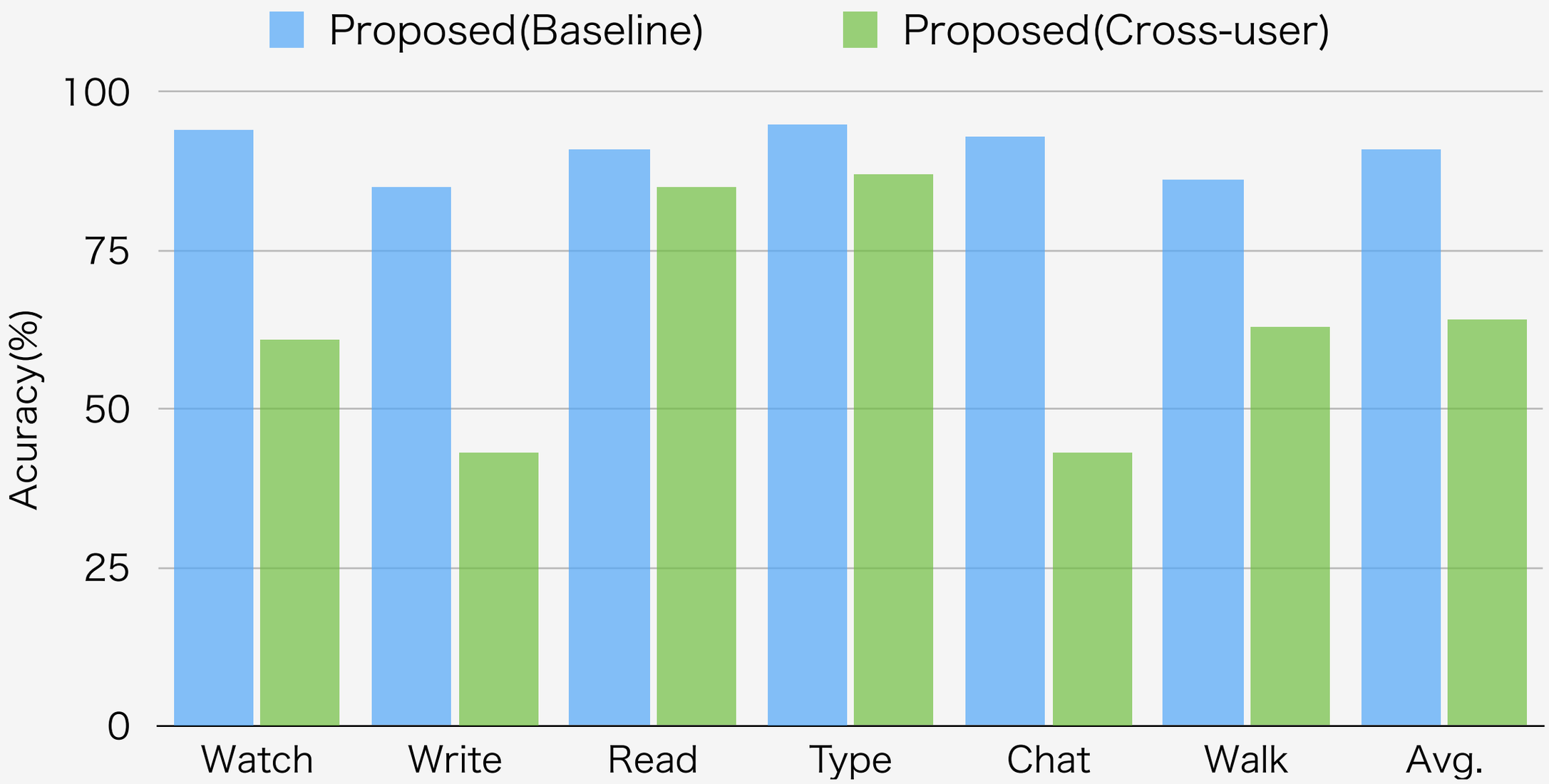
	Wach a video	Write text	Read Text	Type text	Have a chat	Walk
Scene 1						
Scene 2						

×

7 people

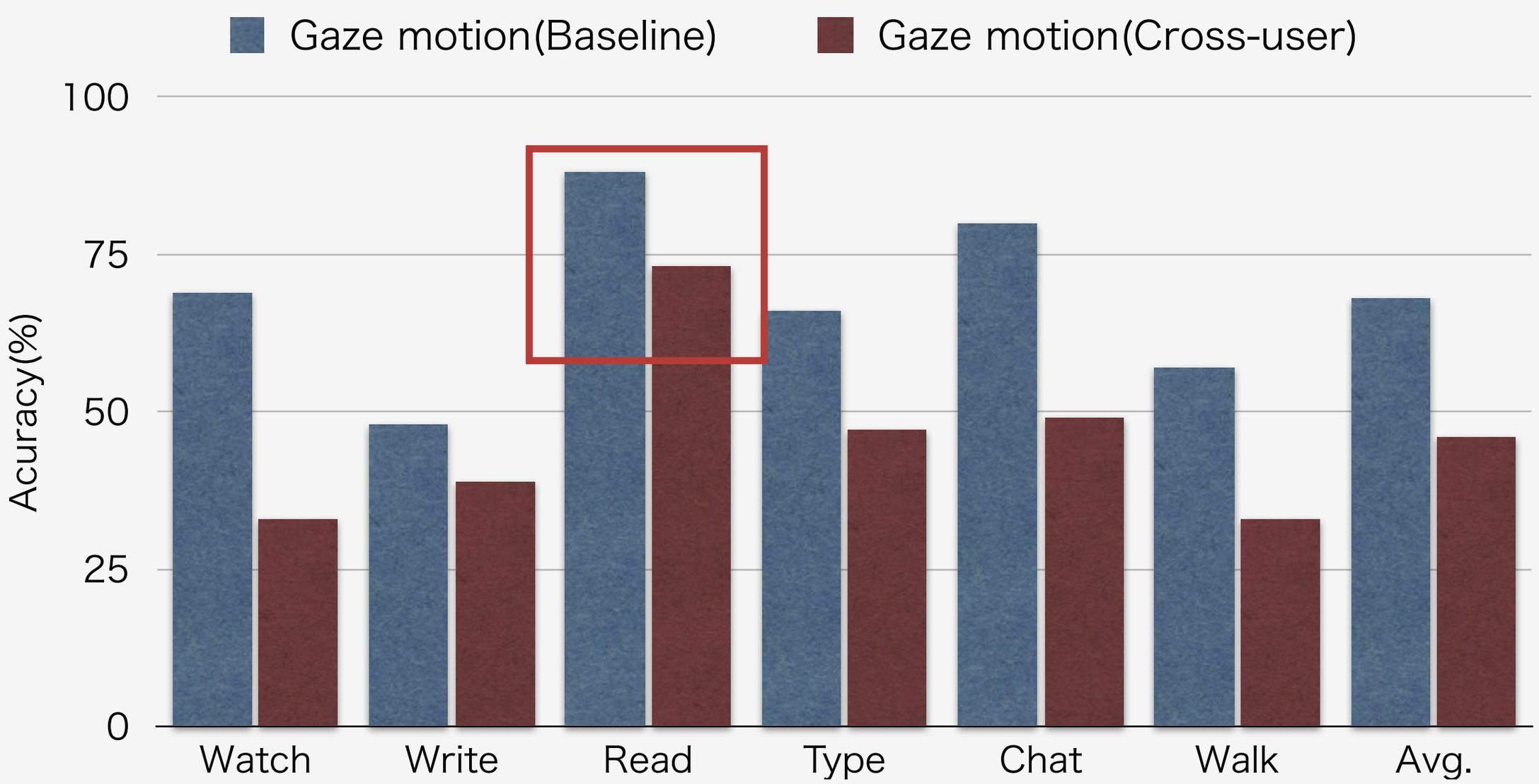
1 person: test　　The rest 6 people: training

Cross-user Experiment



- The recognition rate of Cross-user is lower than Baseline

Cross-user Experiment



- Gaze motions are different between people
- Gaze motions of “Read” activity are similar between different people

Outline

- Introduction
- Proposed Method
- Experiment
- Conclusion

Conclusion

- Combined gaze motion feature and visual feature to recognize daily activities that involve eye movements
- The results from the experiments show that the recognition accuracy is higher when we combine vision-based method and gaze motion-based method

Daily Activity Recognition Combining Gaze Motion and Visual Features

Yuki Shiga, Takumi Toyama, Yuzuko Utsumi,
Andreas Dengel, Koichi Kise

大阪府立大学
OSAKA PREFECTURE UNIVERSITY

Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

Cross-User Experiment