Daily Activity Recognition
Combining Gaze Motion
and Visual Features

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

- 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)

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

Output

Fusion

Classifier

Output

Result
Overview

Eye Tracker

Record Gaze Points and Scene Images

Gaze Motion Feature

Visual Feature

Classifier

Output

Fusion

Classifier

Output

Result
Gaze Motion Feature

- The method proposed by Bulling et al.

Overview

Eye Tracker → Record Gaze Points and Scene Images → Gaze Motion Feature

Classifier → Fusion → Result

Classifier → Fusion → Result
Crop a region around gaze points to remove an irrelevant region
Crop a region around gaze points to remove a irrelevant region
Local Feature Extraction

Extract Local Features (PCA-SIFT) From Each Point

Intrest Points by Dense Sampling
Convert to Global Feature

Learning Image

Test Image

k-means clustering

k centroids
(visual words)

Nearest Neighbor Search
to visual words

Global Feature
Overview

Eye Tracker

Record Gaze Points and Scene Images

Gaze Motion Feature

Classifier

Output

Fusion

Classifier

Result
Classifer

- 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

Read

Write

Type

Probability
Overview

Eye Tracker

Record Gaze Points and Scene Images

Gaze Motion Feature

Visual Feature

Classifier

Output

Fusion

Output

Classifier

Result
Fusion

Read

Probability from gaze motion

Read
Write
Type

Probability from vision
Fusion

Read, Write, Type: Probability from gaze motion

Read, Write, Type: Probability from vision

Average

Read, Write, Type: Combined probability
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

<table>
<thead>
<tr>
<th></th>
<th>Target Objects / Environments</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Same</td>
<td>Same</td>
</tr>
<tr>
<td>Cross-scene</td>
<td>Different</td>
<td>Same</td>
</tr>
<tr>
<td>Cross-user</td>
<td>Same</td>
<td>Different</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Scene 1</th>
<th>Scene 2</th>
<th>Scene 3</th>
<th>Scene 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wach a video</td>
<td>Write text</td>
<td>Read Text</td>
<td>Type text</td>
</tr>
<tr>
<td>Have a chat</td>
<td>Walk</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th></th>
<th>Watch a video</th>
<th>Write text</th>
<th>Read Text</th>
<th>Type text</th>
<th>Have a chat</th>
<th>Walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene 1</td>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
<td><img src="image3.png" alt="Image 3" /></td>
<td><img src="image4.png" alt="Image 4" /></td>
<td><img src="image5.png" alt="Image 5" /></td>
<td><img src="image6.png" alt="Image 6" /></td>
</tr>
<tr>
<td>Scene 2</td>
<td><img src="image7.png" alt="Image 7" /></td>
<td><img src="image8.png" alt="Image 8" /></td>
<td><img src="image9.png" alt="Image 9" /></td>
<td><img src="image10.png" alt="Image 10" /></td>
<td><img src="image11.png" alt="Image 11" /></td>
<td><img src="image12.png" alt="Image 12" /></td>
</tr>
<tr>
<td>Scene 3</td>
<td><img src="image13.png" alt="Image 13" /></td>
<td><img src="image14.png" alt="Image 14" /></td>
<td><img src="image15.png" alt="Image 15" /></td>
<td><img src="image16.png" alt="Image 16" /></td>
<td><img src="image17.png" alt="Image 17" /></td>
<td><img src="image18.png" alt="Image 18" /></td>
</tr>
<tr>
<td>Scene 4</td>
<td><img src="image19.png" alt="Image 19" /></td>
<td><img src="image20.png" alt="Image 20" /></td>
<td><img src="image21.png" alt="Image 21" /></td>
<td><img src="image22.png" alt="Image 22" /></td>
<td><img src="image23.png" alt="Image 23" /></td>
<td><img src="image24.png" alt="Image 24" /></td>
</tr>
</tbody>
</table>

- 3 people
## Cross-scene Experiment

<table>
<thead>
<tr>
<th>Scene</th>
<th>Watch a video</th>
<th>Write text</th>
<th>Read Text</th>
<th>Type text</th>
<th>Have a chat</th>
<th>Walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
<tr>
<td>3</td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
<td><img src="image16" alt="Image" /></td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
</tr>
<tr>
<td>4</td>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" alt="Image" /></td>
<td><img src="image21" alt="Image" /></td>
<td><img src="image22" alt="Image" /></td>
<td><img src="image23" alt="Image" /></td>
<td><img src="image24" alt="Image" /></td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th></th>
<th>Wach a video</th>
<th>Write text</th>
<th>Read Text</th>
<th>Type text</th>
<th>Have a chat</th>
<th>Walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene 1</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td>Scene 2</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
</tbody>
</table>

×

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
Cross-User Experiment

Accuracy (%)

- Watch
- Write
- Read
- Type
- Chat
- Walk
- Avg.

Comparison between Visual (Baseline) and Visual (Cross-user)