Abstract
Recognition of user activities is a key issue for context-aware computing. We present a method for recognition of user daily activities using gaze motion features and image-based visual features. Gaze motion features dominate for inferring the user’s egocentric context whereas image-based visual features dominate for recognition of the environments and the target objects. The experimental results show the fusion of those different type of features improves performance of user daily activity recognition.

Author Keywords
activity recognition; context-awareness; eye tracking; bag-of-features; image recognition;

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction
Context-awareness recently draws public attention in the field of pervasive computing [8]. When a computer knows the context of the user, an adequate service could be supplied to him or her at a proper moment. Museum Guide 2.0 [14] is one of the examples of such a system, which detects the user attention to art
objects in a museum and presents additional information about them at a right moment.

For context-aware systems, recognition of user activities is a key issue. By recognizing which type of activity the user is performing, the inference of the whole user context becomes more reliable. In this paper, we focus on daily activities which could occur in everyday life. Recent developments of wearable computing devices open up the opportunities for weaving technologies into our everyday scenes. In particular, eye tracking devices, which become more easily available and portable in recent years, have great potential for progress of context aware systems in the near future.

Previously, several approaches have been proposed for activity recognition using a mobile eye tracker [3] [7] [12] [6]. These approaches are either based on gaze motion features or visual features from a point of gaze focus. In this paper, we propose a recognition method which combines a gaze motion-based approach and a vision-based approach, inspired by the approaches of previous work. It has been shown that there are some daily activities which could be easily recognized by visual features and some other activities could also be recognized by gaze motion features. For example, when a user writes text on a white paper, “a hand holding a pen” is dominantly in user’s focus (at least in a peripheral region of user’s focus). If we visually recognize “a hand holding a pen” in the user’s focus, it is inferred that the user is doing something with a pen. On the other hand, when we read text, our gaze motion draws characteristic patterns, as the gaze position repeatedly moves from left to right (if it is English text). Keeping these notions in mind, we propose a method combining these two approaches for daily activity recognition.

In this paper, we describe the proposed method and present the experimental results for evaluation of the proposed method. In the experiments, we evaluate whether the combination of two different features can improve the recognition performance in different environments, and whether the proposed method can also be applied to cross-users. The experimental results show the benefit of such a fusion of multiple features.

Related Work
Since the advent of eye tracking technology, a number of researches have contributed to understand the nature of human eye movements. One of the breakthroughs of eye tracking research was the employment of eyes as an interface for computer interaction [2]. Since then, a number of approaches have been proposed for human-computer interaction using eye trackers. Recognition of gaze activities is a relatively new topic. Until few years ago, it was only usable when the user is facing to a computer screen [1] [9].

Nowadays, we can exploit a wearable eye tracker in various daily scenarios where activity recognition in particular becomes a trend. A typical example is the framework presented in [3], which uses gaze motion features obtained from an electrooculography. In [12], Ogaki et al. extended the method by compensating motion features based on camera image tracking. Both approaches focus on the activities in an office working scenario, where visual features are not very relevant.
It is obvious that recent developments of computer vision technologies have also contributed to user activity recognition. Especially, interests on image based activity recognition using an egocentric vision [10] are growing because of the spreading of wearable cameras such as Google Glass. In [7] and [6], eye-gaze and its temporal relationship are also used in order to extract features from the user’s focus. Unlike these previous contributions, our method extracts motion features from gaze movement patterns based on the approach proposed by [3]. The final recognition output is determined by fusion of the image-based visual recognition result and motion-based recognition result.

Furthermore, the activities and actions addressed in the previous work are in most cases only in a particular scenario (e.g., office working, preparing a food, etc.),

---

1 http://www.google.com/glass/start/
where such a fusion of multiple features is not really significant. One of the main differences of our framework to other work is that the activities we focus (refer to Figure 1) can be said relatively generic. When the activity classes become more generic, information from different channels could compensate each recognition task more efficiently.

Proposed Method
Architecture Overview
In Figure 2, we show an overview of the architecture of the proposed method. For eye tracking, we use SMI Eye Tracking Glasses (ETG). From ETG, a scene image as well as the gaze position in the image is streamed continuously. The recognition framework consists of two parts. One is the recognition part using visual features from the scene images and the other is using gaze motion features. The results from both recognition parts are fused after each recognition process to obtain a final recognized activity result.

Eye Tracking and Scene Image
ETG has an outward facing scene camera (1280x960 pixels and 25 fps). Two inward facing eye cameras are also integrated to track the binocular eye movements. Scene images and coordinates of gaze positions are sent from the eye tracking server in real-time.

Visual Feature Extraction and Classification
Similar to the method proposed in [14], we use a local feature-based method for activity recognition, which is broadly used for object recognition. First, a local image working, visual features could be more helpful, depending on the classes.

2 For example, to recognize “copy text” or “read text” could be done only using gaze motion feature in an office working scenario. However, if we want to recognize more generic activity class such as “read text” or “have a chat” apart from office

3 http://www.eyetracking-glasses.com/
of the user’s focus is cropped from a received scene image using a gaze region. Then from the local image (300x300 pixels), local features using PCA-SIFT [13] and Dense-sampling [11] are extracted. We also apply a bag-of-features model [5] to represent the image with a histogram of visual words.

Multiclass Support Vector Machine (SVM) from LibSVM [4] is employed as a classifier. For training, local images from recorded eye tracking videos are collected for each activity. A trained multiclass SVM outputs the probability score of each class \( p_i \):

\[
p_i = P(y = i | x), \quad i = 1, \ldots, k,
\]

where \( y \) is a class label and \( k \) is the number of classes for given test sample (histogram) \( x \).

**Motion Feature Extraction and Classification**

Gaze motion features proposed in [3] and [12] are employed in our framework. When an eye tracking video sequence \( X \) and a current timestamp \( t \) are given, fixations and saccades are detected from the gaze motion sequence of a local time window \( W \), i.e., the gaze data from \( t-W \) to \( t \) in \( X \) is used to extract features for \( t \).

After detecting fixations and saccades, a fixation and saccade sequence is converted to a codebook sequence, which represents the orientations and the lengths of saccades in the given motion sequence\(^4\).

From the codebook sequence, codebook word sequences of 1-gram, 2-gram, 3-gram, and 4-gram are generated. We extract statistical features from each \( n \)-gram sequence: maximum of word \( (n \)-gram) sequence frequency, number of unique \( n \)-grams, average of word frequency, variance of word frequency, and difference of maximum, and minimum of word frequency.

Similar to vision-based method, a multiclass SVM is trained using extracted gaze motion features for each activity.

**Fusion of Classification Results**

With the motion-based classifier, features are extracted from time window \( W \), whereas an image of a single time frame is used in the vision-based classifier. Thus, for time \( t \), the vision-based method extracts features only from time \( t \) and the motion-based method extracts features from sequence from \( t-W \) to \( t \).

As described above, the outputs of the vision-based classifier and the motion-based classifier represent the probabilities of the given samples belonging to a class \( y \). As proposed in [15], we sum the SVM scores for each activity class in order to fuse the recognition results from two classifiers. The final recognition probability of class \( i \) is given as follows:

\[
p_i = \frac{pv_i + pm_i}{2},
\]

where \( pv_i \) and \( pm_i \) are the probability scores from the vision-based classifier and the motion-based classifier, respectively. The final recognized activity class output is the class \( i \) that maximizes the probability score.

\(^4\) For detail, refer to [12]. We only use gaze motion sequences and do not use image tracking using optical flow.
Experiments
To evaluate the proposed method from several perspectives, we conducted three experiments. First, we describe the dataset we used in the experiments.

Dataset
In this paper, six daily activities, which occur in everyday life are selected. Figure 1 shows the six classes and examples of scene image and gaze data.

For each activity class, four different scenes were recorded from a single user. Using this data, we investigate whether the method can correctly recognize the activities even when the target objects or the environments change. For example, the target objects are different in "read" class. In scene 1, the user reads a document printout, whereas the user reads text on a web page in scene 3. If a target object is very different within the class, vision-based method might have a disadvantage; on the other hand, it can be easily recognized if a target object is similar within a class.

We recorded scene image videos and the gaze data while the participants were performing those activities in each scene. For each scene, a video of 120 sec. is recorded. Seven participants took part in the experiments.

Visual words used in the vision-based method were obtained by clustering the extracted local features from the images of training dataset.

Experiment 1: Baseline Experiment
First, we evaluated the performance of the method in the easiest scenario as a baseline. In this experiment, all the recorded video sequences from a single participant were divided into two parts. The first half was used for training and the other half was used for test. That is, the participants and the scenes were the same in training data and the test data. The test set and the training set was then swapped for calculating the average.

Table 1 shows the recognition results. Since the target objects and the environments were the same in the training data and the test data, the recognition results of vision-based method were relatively good. Not surprisingly, the motion-based method works well in "read" class, which might contain characteristic features in gaze motions. For all classes, the fusion results were better than the individual classifiers.

<table>
<thead>
<tr>
<th></th>
<th>Watch</th>
<th>Write</th>
<th>Read</th>
<th>Type</th>
<th>Chat</th>
<th>Walk</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>87</td>
<td>84</td>
<td>95</td>
<td>84</td>
<td>83</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>69</td>
<td>48</td>
<td>88</td>
<td>66</td>
<td>80</td>
<td>57</td>
<td>68</td>
</tr>
<tr>
<td>F</td>
<td>94</td>
<td>85</td>
<td>91</td>
<td>95</td>
<td>93</td>
<td>86</td>
<td>91</td>
</tr>
</tbody>
</table>


Experiment 2: Cross-scene Classification
Secondly, we evaluated the performance when the scenes are different in the training data and the test data. In this experiment, the videos from scene n were left out for test data and the videos from the rest of the
scenes were used for training. The recognition results were calculated based on leave-one-out cross validation. We investigate the recognition performance when the target objects and the environments change.

<table>
<thead>
<tr>
<th></th>
<th>Watch</th>
<th>Write</th>
<th>Read</th>
<th>Type</th>
<th>Chat</th>
<th>Walk</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>80</td>
<td>66</td>
<td>45</td>
<td>91</td>
<td>28</td>
<td>72</td>
<td>64</td>
</tr>
<tr>
<td>M</td>
<td>53</td>
<td>38</td>
<td>73</td>
<td>61</td>
<td>74</td>
<td>44</td>
<td>57</td>
</tr>
<tr>
<td>F</td>
<td>87</td>
<td>60</td>
<td>78</td>
<td>93</td>
<td>68</td>
<td>70</td>
<td>76</td>
</tr>
</tbody>
</table>

Table 2. Recognition accuracy of each class in Experiment 2: Cross-scene classification.

Table 2 shows the recognition results of this experiment. It is clearly shown that the task becomes more difficult than Experiment 1 since the overall recognition accuracy dropped. Especially, the accuracy of the vision-based method of "have a chat" class decreased drastically. A main reason for this was that many participants also looked background objects frequently during the chats. This results imply that when we have a conversation, we do not always look at the other’s face. It was surprising that the vision-based method achieved good recognition accuracy for "watch a video" class even though the content of the video was different between the training and the test. This was because the screen used in the experiment was visually recognized. Most importantly, the fusion could also improve the individual recognition performance similar to Experiment 1. The fusion results show that each method could compensate for the shortcomings of the other, such as "write", "read", and "chat" classes.

Experiment 3: Cross-user Classification

Lastly, we evaluated the performance when the user is different in the training data and test data. Generally speaking, gaze motions might be different between the individuals but visual features should be similar within a class. In this experiment, we investigate to what degree a trained classifier with other users can be applied to another user. A single participant was left out for test and the other participants were used for training the classifier. Similar to the previous one, leave-one-out cross validation was applied in this experiment. In this experiment, the recorded data from scene 1 and scene 2 is used.

<table>
<thead>
<tr>
<th></th>
<th>Watch</th>
<th>Write</th>
<th>Read</th>
<th>Type</th>
<th>Chat</th>
<th>Walk</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>69</td>
<td>33</td>
<td>79</td>
<td>83</td>
<td>34</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>M</td>
<td>33</td>
<td>39</td>
<td>73</td>
<td>47</td>
<td>49</td>
<td>33</td>
<td>46</td>
</tr>
<tr>
<td>F</td>
<td>61</td>
<td>43</td>
<td>85</td>
<td>87</td>
<td>43</td>
<td>63</td>
<td>64</td>
</tr>
</tbody>
</table>

Table 3. Recognition accuracy of each class in Experiment 3: Cross-user classification.

Table 3 shows the recognition results of this experiment. Compared to the previous experiments, the overall performance dropped again. However, as predicted, the vision-based method was only slightly worse. Similarly, the results also show that the gaze motions of some activities (such as "watch" and "walk") could be different between the users even if they are was done randomly, in order to reduce the cost of testing all combinations.

Note that there is no relation between the activities within the scene. Even if an activity from "read" and an activity from "watch" is grouped in scene 1, it does not necessarily mean that these activities have any commonality. This grouping by scene

---

WORKSHOP: PETMEI
doing the same activity. However, in other activity classes such as “read” class, the motion-based classifier could also work well, which implies that the gaze motions of many participants were similar during a reading activity.

Discussion
The experimental results showed that the recognition performance could be improved by combining vision-based approach and motion-based approach in general. In particular, it was observed that one classifier could compensate shortcomings of the other classifier in many cases, especially when the target objects and the environments are different within an activity class. For some activities, a trained classifier with other users could perform well, although it is not always the case.

In addition to the improvement of performance, the proposed method has another benefit. When applying the activity recognition framework to real applications, the vision-based classifier could also provide an inference of “what” is in user’s focus. Not only by recognizing the activity the user is performing, but also by recognizing what is in the user attention, the application could become more “context-aware”. Although the proposed framework is not yet able to output what the object is, it can be extended easily with the same framework.

Conclusion
In this paper, we proposed a method for recognition of daily activities combining gaze motion and visual features. The experimental results showed that the proposed method could improve the performance of a classifier using a single feature. The method opens up the possibilities of combining features from multiple resource for daily activity classes.

References


