# A classification method of cooking operations based on eye movement patterns

Hiroya Inoue* Nagoya University
Takatsugu Hirayama† Nagoya University
Keisuke Doman Chukyo University
Yasutomo Kawanishi Nagoya University
Ichiro Ide ‡ Nagoya University
Daisuke Deguchi Nagoya University
Hiroshi Murase Nagoya University

## Abstract

We are developing a cooking support system that coaches beginners. In this work, we focus on eye movement patterns while cooking meals because gaze dynamics include important information for understanding human behavior. The system first needs to classify typical cooking operations. In this paper, we propose a gaze-based classification method and evaluate whether or not the eye movement patterns have a potential to classify the cooking operations. We improve the conventional $N$-gram model of eye movement patterns, which was designed to be applied for recognition of office work. Conventionally, only relative movement from the previous frame was used as a feature. However, since in cooking, users pay attention to cooking ingredients and equipments, we consider fixation as a component of the $N$-gram. We also consider eye blinks, which is related to the cognitive state. Compared to the conventional method, instead of focusing on statistical features, we consider the ordinal relations of fixation, blink, and the relative movement. The proposed method estimates the likelihood of the cooking operations by Support Vector Regression (SVR) using frequency histograms of $N$-grams as explanatory variables.

**Keywords:** cooking operations, gaze analysis, eye movement pattern, fixation, blink, $N$-gram, SVR

**Concepts:** •**Computing methodologies** → **Activity recognition and understanding;**

## 1 Introduction

It is no doubt that cooking delicious meals enriches our daily life. In recent years, various services that support us cook meals have emerged. In particular, introducing information technologies to cooking support systems is an effective approach for helping the beginners learn cooking techniques. For example, the VideoCooKing Interface [Doman et al. 2011] provides a short video segment corresponding to the procedure in a recipe by referring to a video database indexed by the pair of a cooking operation and an ingredient. Most existing support systems expect the user to proactively request for information. However, we consider that an interactive system that adaptively supports according to the user's state is more effective for beginners. To realize this, the system needs to understand what s/he is doing or what s/he is planning to do next.

*e-mail: inoueh@murase.m.is.nagoya-u.ac.jp
†e-mail: hirayama@is.nagoya-u.ac.jp
‡e-mail: ide@is.nagoya-u.ac.jp


ETRA 2016, March 14 - 17, 2016, Charleston, SC, USA
ISBN: 978-1-4503-4125-7/16/03
DOI: http://dx.doi.org/10.1145/2857491.2857500

(a) Example of "Cut"

(b) Example of "Mix"

**Figure 1:** *Example of the difference in eye movement between cooking operations. The red dot and the red line represent the current gaze location and the trajectory of gaze locations, respectively.*

Conventional methods that recognize cooking operations are based on visual features extracted from the images taken by a camera fixed above the kitchen [Hayashi et al. 2013][Matsumura et al. 2015]. However, image-based cooking operation recognition is significantly affected by variations of environmental factors such as the difference on lighting conditions and appearance of cookwares.

In this work, we assume that human behaviors are generated through the process of cognition, judgment, and actuation. Considering that visual attention that is closely linked to cognition while cooking meals is important, we analyze the difference of eye movements in the cooking operations as shown in Figure 1. Generally speaking, the eye movements reflect the internal state of humans and the field-of-view frequently includes visual information related to human actions [Li et al. 2013]. Since the evolution of information technology has produced high performance, compact, and inexpensive wearable sensors for measuring both the eye movements and the first person view, it has become easy for ordinary people to use them in everyday life. Understanding cooking operations based on the analysis of eye movements, we can choose important scenes for summarizing cooking videos or thumbnail images for authoring cooking recipes from the first-person view. Also, it can encourage us to understand tips of cooking operations in terms of cognition and judgment by analyzing the difference between skilled users and beginners. Furthermore, we can segment cooking operations by using eye movement which is difficult by image-based methods.

The goal of this paper is therefore to extract eye movement patterns featuring cooking operations. We attempt to classify the cooking operations based on the analysis of eye movement patterns as a preliminary step for such understanding. The pattern analyzed in this paper is the direction and the distance of the relative transition of the eye movement from the previous frame, for example (right-long, right-short, down-long). Bulling et al. succeeded in classifying human behaviors performed at an office desk based on eye movement patterns [Bulling et al. 2011]. Also, Ogaki et al. demonstrated that the joint use of eye movement which reflects subtle change of attentional direction and head motion which reflects larger change, improved the deskwork classification [Ogaki et al. 2012]. However, behavioral characteristics between cooking and deskworking seem to be different. For instance, gaze location while cooking is usually fixed at a cooking ingredient or an equipment, whereas the gaze location while deskworking is always moving on the monitor.

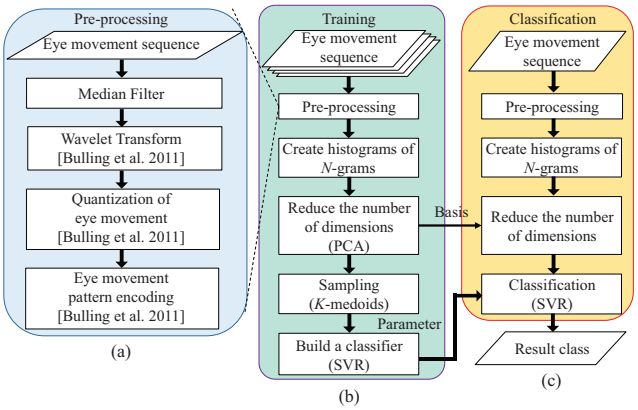


**Figure 2:** *Process flow of classification of cooking operations.*

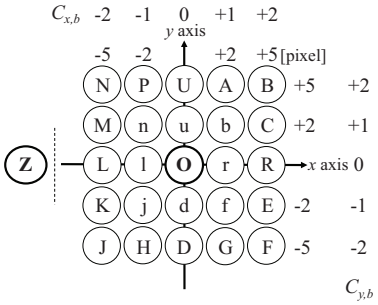


**Figure 3:** *Characters used to encode relative eye movements by direction and distance.*

## 2 Classification method for cooking operations

Figure 2 is a flowchart of the proposed method that first converts eye movement data to character sequences in a pre-processing phase and then calculates histograms from $N$-grams of the character sequence. The histograms are combined to a feature vector to classify the cooking operations. Figure 2(a) shows the detailed procedure in the pre-processing phase. Since the measured eye movement data include noise, we apply a median filter independently to the horizontal ($x$) and the vertical ($y$) components of the data sequence. We then apply CWT-SD (Continuous Wavelet Transform for Saccade Detection) to the filtered data in the same manner as in the conventional method [Bulling et al. 2011] in order to quantize the relative eye movement between frames to $C = \{C_{x,b}, C_{y,b}\}_{b=1}^{T}$. Here, $C_{x,b}$ is the quantized relative movement of $x$ coordinate from frame $b-1$ to $b$. The scale parameter $\alpha$ of the transformation depends on the sampling interval $T$ of the eye movement data ($T = 100$ ms sets $\alpha$ to 6 in the following experiment).

$$C_{x,b} = \frac{1}{\sqrt{\alpha}} \int \psi\Big(\frac{t-b}{\alpha}\Big) x_t \, dt \tag{1}$$

$$\psi(\beta) = \begin{cases} 1 & (0 \le \beta < \frac{1}{2}) \\ -1 & (\frac{1}{2} \le \beta < 1) \end{cases} \tag{2}$$

This process is also applied to the $y$ coordinate. After quantizing the data to 5 bits by thresholding ($\pm H$ and $\pm L$), we encode the eye movement data to a character sequence by integrating $C_{x,b}$ and $C_{y,b}$.

Figure 3 shows the characters used to encode the relative eye movements by distance and direction. The upper and the lower case characters represent the eye movements with longer and shorter distances, respectively. We assume that, for instance, long-distance movements appear while mixing chopped ingredients to follow them with the eyes, while short-distance movements or no movement appear while cutting an ingredient to focus on it. Although the users would often keep their attention to static ingredients, the conventional encoding method [Bulling et al. 2011] does not assign any character to no movement. Thus, we introduce character "O" to the origin in Figure 3, that represents no movement between frames. We also introduce the character "Z" to represent frames during eye blinks. We assume that eye blinks depend on cooking operations because people blink depending on the state of cognition and visual environment [Schleicher et al. 2008]. Although Bulling et al. adopted the rate and duration of fixations and blinks as features, we include each frame during the blink, i.e. character "Z" to the character sequence to analyze the ordinal relation among the relative movement, no movement, and the blink. In summary, in the proposed method, the eye movement sequence is encoded by 26 kinds of characters, two more than in the conventional method. We will describe how to detect eye blinks later in Section 3.

We assume that local patterns in the character sequence depend on the cooking operations. In this work, we put windows with short time length on the character sequence so that each window should include one operation to create training and test sets; windows with a length of 900 frames (15 sec.) following the conventional method [Bulling et al. 2011] are put at 60 frames interval. If the operation changes to another operation within 60 frames following a window, we expand the window to include the frames while the operation continues. Next, frequency histograms of $N$-grams ($N$ = 1, 2,..., $n$) of the characters are created in each window. Bulling et al. extracted statistical features such as mean, variance, and maximum of the frequency from the histograms [Bulling et al. 2011]. Instead of such features, we generate a feature vector by combining all the histograms of $N$-grams to analyze the eye movement patterns to extract the behavioristic characteristics sufficiently. Since the dimension of feature vector will become very high, we apply PCA (Principle Component Analysis) to reduce it.

We finally construct an SVR[1] (Support Vector Regression) [Collobert and Bengio 2001] model that maximizes the likelihood of the feature vectors extracted from the training data for a class of cooking operation and minimizes that for the other classes. We regard this model as a one-against-all classifier that outputs a score to estimate whether or not the test data belongs to the class in the range of $-1$ to 1. In the classification phase shown in Figure 2 (c), when the score is 0 or higher, the test data is classified as the target operation. Here, since the number of negative samples is larger than that of positive samples, the trained classifier might give the negative class an advantage. Therefore, we thin out the negative samples by applying the $k$-medoids clustering [Vinod 1969] depending on the ratio of positive samples to negative samples. The value of $k$ is determined as the number of positive samples. We adopt the nearest training data from each centroid as a negative sample.

## 3 Experiments

We conducted an experiment to classify cooking operations in seven first-person view videos including two cooking recipes by four subjects to verify the effectiveness and robustness of the proposed method. In four of them, four subjects cooked a hamburger steak that included three kinds of cooking operations: "Cut," "Mix," and "Wait". The other three subjects cooked a potato salad that in-

[1]We used epsilon-SVR in "lib-SVM."

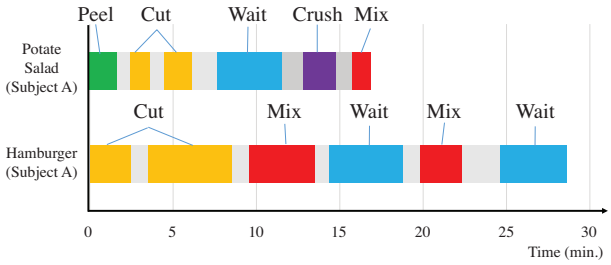


**Figure 4:** *Examples of cooking operations in our dataset.*

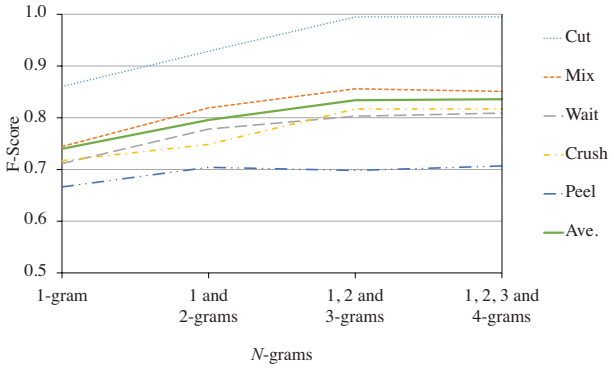


**Figure 5:** *Relationship between accuracy and used $N$-grams.*

cluded five kinds of cooking operations: "Peel," "Cut," "Crush," "Mix," and "Wait". The definition of each cooking operation follows the common taxonomy [Hayashi et al. 2013]. The subjects cooked according to the procedure in each recipe. Figure 4 shows the sequence of cooking operations in each recipe. Regarding the flexibility of the experimental setup, we allowed the subjects to change the layout of cookwares. We employed EMR-9 manufactured by nac Image Technology, Inc. to measure gaze location on the first-person-view video as eye movement data [NAC ]. Its measurable view angle was $\pm 40^\circ$ in horizontal and $\pm 20^\circ$ in vertical directions, and its sampling frequency was 60 Hz. The resolution of gaze location was $0.1^\circ$ in horizontal and $0.1^\circ$ in vertical directions, and the resolution of the video was 640 (H) $\times$ 480 (V) pixels. Regarding the calibration, nine markers were placed on a plane including the chopping board and the subjects were asked to turn their gaze at each marker in order. We converted the gaze location data into the character sequence as described in Section 2. As a result, we obtained 7,880 samples (character sequences). Each sample corresponded to one of the five kinds of cooking operations. We applied leave-one-video-out cross-validation for the evaluation. We regard frames where EMR-9 failed to measure gaze location as blinks and assign the character "Z" to each of them. To apply CWT-SD to the whole gaze location sequence in the preprocessing phase, we applied linear interpolation to the unmeasured frames.

### 3.1 Pre-experiment

The encoding process of the gaze location data to the character sequence involves two undefined thresholds: $L$ and $H$. We attempt to set them appropriately to distinguish between short and long-distance eye movements, in other words, the lower and upper case characters in Figure 3, so that we have the potential for classifying the cooking operations based on the eye movement patterns. Threshold $L$ was determined to detect frames during fixation defined as stable gaze within one degree over 100 ms [Irwin 2004]. Here, one degree angle of view equals to 12 pixels on the image coordinate of EMR-9 and 100 ms equals to 6 frames in the video taken by EMR-9. Therefore, threshold $L$ was set to 2 pixels. On the other hand, threshold $H$ was determined experimentally because no hypothesis exists to set an appropriate value. As a result of experiment in which we used all the datasets, the maximum F-score of classification was obtained when $H = 5$ pixels.

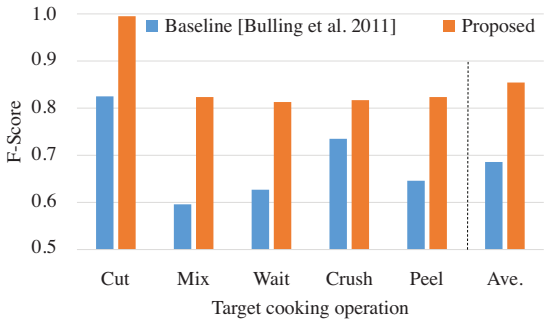


**Figure 6:** *Accuracy comparison with the conventional method in classification of cooking operations.*

Since the proposed method uses the combined histogram of $N$-grams ($N$ = 1, 2, ..., $n$) as a feature vector, we also need to determine the maximum value $n$ appropriately. Figure 5 shows the relationship between F-score and $N$-grams used for each cooking operation. As $n$ increases, F-score improved. However, F-score was hardly improved for $n$ more than 4 even though the classification used a higher dimensional feature vector, and thus required higher computational cost. We therefore used the combined histogram of $N$-grams up to 3-gram in the following experiments.

### 3.2 Comparison with the conventional method

We employ a baseline method based on an $N$-gram model of the eye movement pattern [Bulling et al. 2011] which works well for recognizing operations performed at an office desk. In this method, an $N$-gram wordbook consists of 24 types of characters without "O" that represents no movement and "Z" that represents blink, and the following five statistical features are extracted from the wordbook: (1) max-count, (2) average-count, (3) wordbook size, (4) variance of counts, and (5) difference between maximum and minimum counts. In fact, the features are extracted from each of four wordbooks up to 4-gram.

Figure 6 shows the F-score of classification yielded from the proposed and the baseline methods for each cooking behavior. We confirmed that the proposed method was superior to the baseline method for all cooking operations. The proposed method marked a higher average F-score of 0.854. We confirmed whether or not there were individual differences in accuracy. The result suggests that eye movement depends on basic operations but not on subjects.

### 3.3 Effects of two additional eye movement characters

We conducted an experiment to verify the effectiveness of the additional characters "O" and "Z". Figure 7 shows the F-score for (1) the baseline method [Bulling et al. 2011], (2) adopting the combined histogram of $N$-grams with 24 conventional characters as a feature vector instead of statistical features, (3) adding "O," (4) adding "Z," and (5) adding "O" and "Z," to (2). (5) is the proposed method. As we can see from the average the F-score, each additional character contributed to achieve better classification. Especially in "Cut," "Crush", and "Peel," "O" and "Z" contributed to the improvement of the F-score. However, F-score of the proposed method for "Mix" and "Wait" was not the highest. We cite as a reason that the number of training data for obtaining distinctive patterns was not sufficient, due to increasing feature dimensions.

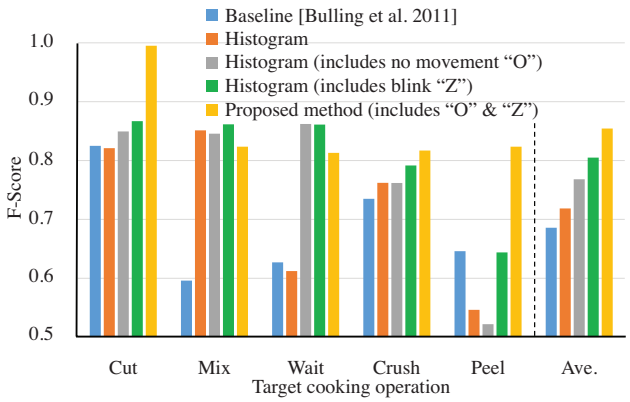


**Figure 7:** *Effects of the additional eye movement characters on classification of cooking operations.*

### 3.4 Multi-class classification

To estimate the cooking operation for each window, we need to choose a class with the maximum classification score by applying all the one-against-all classifiers to the feature vector. This means multi-class classification. Figure 8 shows the confusion matrix normalized across ground-truth rows, precision and recall of the classification result. The proposed method could estimate the correct class for more than 50% for "Cut," "Mix," and "Wait" classes. The proposed method obtained high precision except for "Mix" class. The recall for "Mix" class was very high, while the precision was low. This is probably because the eye movement patterns while mixing is common to other operations, too. Moreover, the estimated classes often changed window by window. We need to consider the temporal consistency of cooking operation across a longer interval.

Predicted class

| Actual class | Cut | Mix | Wait | Crush | Peel | Recall |
|---|---|---|---|---|---|---|
| Cut | 0.67 | 0.31 | 0.02 | 0.00 | 0.00 | 0.67 |
| Mix | 0.01 | 0.98 | 0.01 | 0.00 | 0.00 | 0.98 |
| Wait | 0.04 | 0.45 | 0.51 | 0.00 | 0.00 | 0.51 |
| Crush | 0.05 | 0.53 | 0.00 | 0.41 | 0.00 | 0.41 |
| Peel | 0.17 | 0.53 | 0.06 | 0.01 | 0.23 | 0.23 |
| Precision | 0.88 | 0.47 | 0.90 | 0.99 | 1.00 | |

**Figure 8:** *Result of multi-class classification.*

## 4 Conclusion

In this paper, we proposed a cooking operation classification method based on the analysis of eye movement patterns. We obtained higher accuracy through the one-against-all classification experiment. However, the proposed method could not achieve accurate multi-class classification. In future work, we need to compare and combine the proposed method with an image-based method that extracts visual features from the first-person-view video. Also, we are planning to apply feature selection methods such as Random Forests for classification instead of SVR, in order to analyze the relation between the selected useful features and psychological knowledge as in [Land and Hayhoe 2001]. To conduct a more detailed and comprehensive analysis of the eye movement patterns, we are planning to apply Latent Dirichlet Allocation (LDA) to a larger scale dataset, which is useful for classifying usual activities in daily life [Steil and Bulling 2015]. The patterns extracted by LDA will encourage us to analyze the difference in cooking operations between skilled users and beginners.

## Acknowledgements

Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research.

## References

BULLING, A., WARD, J., GELLERSEN, H., AND TROSTER, G. 2011. Eye movement analysis for activity recognition using electrooculography. *Proc. IEEE Trans. on Pattern Analysis and Machine Intelligence 33*, 4 (February), 741–751.

COLLOBERT, R., AND BENGIO, S. 2001. Support vector machines for large-scale regression problems. *J. Machine Learning Research 1* (February), 143–160.

DOMAN, K., KUAI, C., TAKAHASHI, T., IDE, I., AND MURASE, H. 2011. Video CooKing: Towards the synthesis of multimedia cooking recipes. In *Proc. 17th Int. Conf. on Multimedia Modeling*, 135–145.

HAYASHI, Y., DOMAN, K., IDE, I., DEGUCHI, D., AND MURASE, H. 2013. Automatic authoring of domestic cooking video based on the description of cooking instructions. In *Proc. 5th Int. Workshop on Multimedia for Cooking and Eating Activities*, 21–26.

IRWIN, D. 2004. Fixation location and fixation duration as indices of cognitive processing. In *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, Psychological Press, 105–134.

LAND, M., AND HAYHOE, M. 2001. In what ways do eye movements contribute to everyday activities? *Vision Research 41*, 25 (November), 3559–3565.

LI, Y., FATHI, A., AND REHG, J. 2013. Learning to predict gaze in egocentric video. In *Proc. 2013 IEEE Int. Conf. on Computer Vision*, 3216–3223.

MATSUMURA, Y., HASHIMOTO, A., MORI, S., MUKUNOKI, M., AND MINOH, M. 2015. Clustering scenes in cooking video guided by object access. In *Proc. 7th Int. Workshop on Multimedia for Cooking and Eating Activities*, 1–6.

NAC. *http://www.nacinc.com/datasheets/archive/EMR9-DataSheet-June-09.pdf*. Accessed Jan. 2016.

OGAKI, K., KITANI, K., AND SUGANO, Y. 2012. Coupling eye-motion and ego-motion features for first-person activity recognition. In *Proc. IEEE Workshop on Egocentric Vision in Conjunction with CVPR2012*, 1–7.

SCHLEICHER, R., GALLEY, N., BRIEST, S., AND GALLEY, L. 2008. Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired? *Ergonomics 51*, 7 (July), 982–1010.

STEIL, J., AND BULLING, A. 2015. Discovery of everyday human activities from long-term visual behavior using topic models. In *Proc. 2015 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing*, 75–85.

VINOD, D. 1969. Integer programming and the theory of grouping. *J. American Statistical Assoc. 64*, 326 (June), 506–519.