Information Engineering Express
International Institute of Applied Informatics
2019, Vol. 5, No. 1, 27 – 36

Hand Detection in Egocentric Video and Investigation Toward Fine-Grained Cooking Activities Recognition

Katsufumi Inoue *, Misa Ono *, Michifumi Yoshioka *

Abstract

The analysis of egocentric videos is currently a hot topic in computer vision. In this paper, we focus on cooking activities recognition in egocentric videos. To recognize cooking activities automatically and precisely, we must solve the problems of detecting hand region in egocentric videos, representing hand motion, and classifying the cooking activities. In this research, to solve these problems, we propose a new cooking activities recognition method in egocentric videos. The characteristic points are 1) hand regions are accurately detected in a cluttered background by using color, texture, and location information, 2) temporal hand features are extracted from sequential frame images with a thinning algorithm, 3) a fully-connected multi-layer neural network is utilized to recognize the activities from the extracted features. Toward our goal of fine-grained cooking activities recognition, we in-vestigated the performance of our method with our benchmark, including 12 fine-grained cooking activities in five coarse categories. The experimental results show that our method allows us to recognize cooking activities with an accuracy of 45.2%.

Keywords: Egocentric Video, Fine-grained Cooking Activities Recognition, Fully Con-nected Multi-layer Neural Network, Hand Detection

1 Introduction

Advances in wearable camera such as Go-Pros have sparked great interest in egocentric video analysis in computer vision. Compared with third-person cameras such as surveillance cameras, first-person cameras such as head-mounted cameras have the advantages that they are hands-free and can capture objects touched or manipulated by the camera wearer with almost no occlusions. Consequently, the analysis of egocentric videos promises to be useful for various applications, such as skills transfer of traditional techniques and for learning recipes. In this research, we take the first step toward fulfill these expectations by focusing on cooking activities recognition in egocentric videos.

Related work [1][2] has considered which objects manipulated by the camera wearer may be related to the different cooking activities and often classifies the cooking activities by recognizing the objects. However, due to cluttered backgrounds, object recognition in

* Osaka Prefecture University, Osaka, Japan

egocentric videos is difficult. In contrast to this approach, we focus on repetitive motions, such as cutting vegetables and peeling fruits.

For activities recognition based on motions, we need to detect hand regions accurately in egocentric videos, represent hand motion, and classify the cooking activities based on hand motion features. In this paper, we propose a novel hand detection and cooking activities recognition method for egocentric videos. In our method, to detect hand regions, we exploit the latest superpixel segmentation algorithm and hand region estimation based on the hand location and skin color information. To represent hand motion, we extract the temporal hand motion features by representing hand shapes with a thinning algorithm. To classify the cooking activities, we utilize multi-layer neural network to recognize cooking activities from the extracted temporal features.

Moreover, to understand the cooking scenes in egocentric videos, we need to recognize the fine-grained cooking activities. For example, in the "Cutting" activity, we must distinguish whether the camera wearer is cutting a vegetable into thin strips or chopping it into small pieces. In the first step toward fine-grained recognition, we investigate the perfor-mance of our method with five coarse cooking activities, including a total of 12 fine-grained activities. This paper extends the previous work done by Inoue et al. [3].

2 Related Work

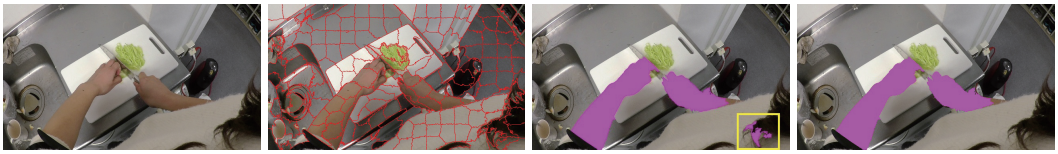
2.1 Hand Region Detection

Hand detection is an important processing method for recognizing cooking activities in egocentric videos. The most traditional approach is based on skin color [4]. Although this approach is simple and effective detecting hand regions, skin-colored objects in the background affect the detection performance. Another approach is based on appearance models generated from images [5] or 3D models [6]. In these methods, once the hand models are generated, we can detect the region effectively. However, many appearance models must be generated and a similar model must be found in a large database for the input image.

In contrast to these approaches, a pixel-level hand region detection method has bee pro-posed that uses color and texture information [7][8]. After superpixel segmentation of each frame image in the video, color information, such as RGB, HSV, and LAB information, as well as texture information extracted by Gabor filter and local descriptors, such as SIFT [9], are utilized to detect hand regions precisely. Our approach is based on this type of approach; however, inspired by clothing parsing tasks [10][11], we exploit the location information of hand region in addition to the color and texture information because hand gestures for cooking activities tend to appear in similar locations in egocentric videos.

2.2 Hand Motion Representation

Egocentric videos undergo large egomotion; therefore, hand motion representation is important for activities recognition in egocentric videos. Li et al. [12] extract hand motion by averaging optical flow vectors in the detected hand regions. In addition, Li et al. [13] exploit local descriptors from Dense Trajectories [14] for hand motion representation. Although such representations have led to improvements in whole-body activities recognition [15], hand activities recognition in egocentric videos is still challenging task due to the lack of hand information contained in local descriptors. Compared with the above approaches, we propose a simple, effective hand motion representation method



(a) Original input image (b) Superpixel segmentation (c) Hand region candidates (d) Detection results

Figure 1: Overview of hand region detection. (a) Original input image. (b) Image seg-mented with superpixels by using LSC. Red lines show superpixel boundaries. (c) Hand region estimation results. Purple regions indicate hand region candidates. (d) Hand region detection results after noise removal.

using hand shape information extracted from the sequential frame images with a thinning algorithm [16].

2.3 Activities Recognition

Similar to third-person activities recognition, the basic approach to recognizing the first-person activities is also to analyze features extracted from successive frames in the video sequentially. Fathi et al. [2] employ a hierarchal graph model and Soran et al. [17] utilize HMM model to recognize first-person activities. Li et al. [13] utilize gaze information and combine motion and object features for recognition. Zhou and Berg [18] compare various kinds of classifiers for predicting of first-person activities with extracted motion features, including a simple nearest neighbor search, nearest neighbor search using Dynamic Time Warping [19], a linear regression model, a linear SVM model, and fully connected neural network proposed in [20]. They found that the fully connected neural network achieved highest recognition performance. Inspired by this finding, we also utilize a fully connected multi layer neural network (FCMLNN).

3 Proposed Method

In this section, we explain the processes of our cooking activities recognition method. Our method consists of hand region detection, hand motion representation, and activities recognition.

3.1 Hand Region Detection Method

Figure 1 shows an overview of hand region detection processes. Cluttered backgrounds make the information noisy, and it is difficult to detect hand region precisely. To reduce the effect, we utilize superpixel-level label estimation for hand region detection, similar to [7][8]. We employ Linear Spectral Clustering (LSC) [21], which is a state-of-the-art super-pixel segmentation algorithm, to detect hand region accurately because the segmentation algorithm requires high boundary recall. Figure 1 (b) shows an example of a superpixel boundary image obtained with LSC.

After superpixel segmentation, we extract the color and texture information from each superpixel. We employ RGB, HS in HSV color space, and AB in LAB color space as color information. We do not utilize V and L information to reduce the effect of illumination conditions. In this paper, we describe the RGB, HS, AB features as C_{RGB}, C_{HS} and C_{AB}, respectively.

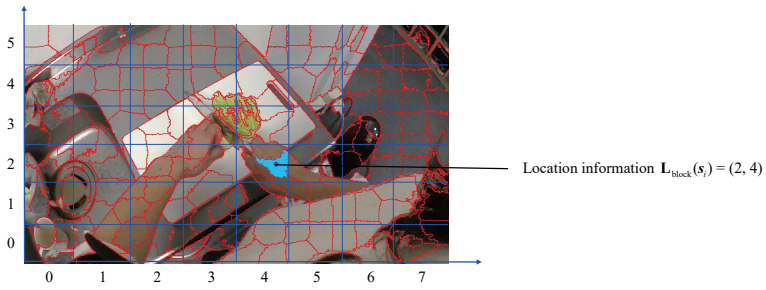


Figure 2: Overview of location indication vector $\mathrm{L}_{\mathrm{block}}$. We divide the input image into 6 × 8 blocks in our experiments. We detect a block which the centroid of superpixel s_i belongs. For example, the location indication vector, $\mathrm{L}_{\mathrm{block}}$ of superpixel s_i (aqua region) is (2, 4).

For texture information extraction, we utilize a Gabor filter, and we describe Gabor feature as $\mathrm{T}_{\mathrm{Gabor}}$. In addition, inspired by clothing parsing tasks [10][11], we also extract location information of the hand region from each superpixel. The location information is useful for detecting hand region because hand motions for cooking activities tend to appear in a similar location in egocentric videos. Figure 2 shows an overview of the extraction process. First, we divide up the input image and we calculate a location indication vector $\mathrm{L}_{\mathrm{block}}$ for each superpixel, which indicates the block to which the centroid of each superpixel belongs. In Fig. 2, $\mathrm{L}_{\mathrm{block}}$ of aqua superpixel s_i is (2, 4).

Next, by concatenating these pieces of information, we estimate whether each super-pixel belongs to the hand region or background with an SVM with RBF kernel and we collect hand region candidate superpixels. However, in Fig. 1 (c), background superpixels are still included in the candidates. To exclude these noisy superpixels, we integrate the adjacent candidate superpixels and detect two integrated superpixels as hand regions. Figure 1 (d) shows an example of detected hand regions.

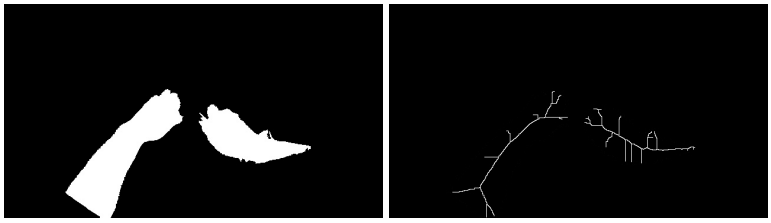
3.2 Hand Motion Representation Method

In this section, we explain how we represent hand motion. Our method exploits sequential hand shape change information.

First, to extract the hand shape information, we utilize the hand region mask image obtained by the hand region detection process and obtain the skeleton information for hand region by using a thinning algorithm [16] as shown in Fig. 3. Next we extract a HLAC [22] feature with dimensions of 25 from this skeleton image. By applying these processes to n successive frame images, we obtain n HLAC features and concatenate them. We utilize the concatenated feature as a hand motion feature.

3.3 Activities Recognition Method

In this section, we explain how we recognize cooking activities. Inspired by [18], we introduce a new FCMLNN as illustrated in Fig. 4. The main difference between the method employed in [18] and our method is the number of network layers and the activation function. In [18], they utilize a three-layer FCMLNN and Rectified Linear Unit (ReLU) [23] as activation function in the first and second layer of FCMLNN. In contrast, we utilize a four-layer FCMLNN. Additionally, from first to third layers of our FCMLNN, we utilize



(a) Hand region mask (b) Skeleton of hand region

Figure 3: Example of skeleton information obtained with thinning algorithm

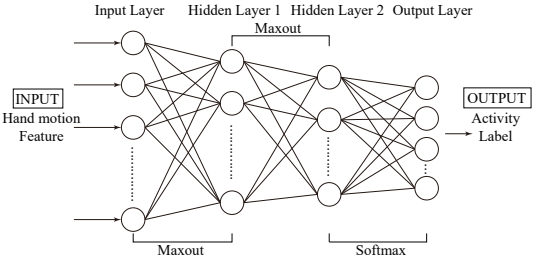


Figure 4: Network structure of our method.

maxout [24] as activation function, which has been found empirically to recognize cooking activities effectively [3]. The activation function of final layer is softmax. The input of our FCMLNN is the extracted hand motion feature mentioned in Sec. 3.2 and the output is the cooking activity label.

4 Experiments

We evaluated our method for cooking activities in egocentric videos with four subjects. In the experiments, toward fine-grained recognition, we captured 12 cooking activities (in detail, see Table 1) to evaluate our method. . The four subjects wore a wearable camera (HX-A500, Panasonic) on the left temporal region of the head near the left ear. Each activity was captured with 30 fps and all videos were composed of 300 frame images with a resolution of 640 × 360. We collected a total of 48 egocentric videos. We focused on deter-mining the effectiveness of our proposed hand region detection method and its performance for fine-grained cooking activities recognition.

4.1 Experiments of Hand Region Detection

We evaluated the effectiveness of our proposed hand region detection method. We utilized 36 videos (3 subjects × 12 activities) for training and other 12 videos were utilized for testing. In addition, from each video, we randomly extracted 10 frame images; we employed 360 frame images for training and 120 frame images for testing. We conducted hand re-gion detection experiments with three different combinations of color, texture, and location features.

To compare the hand regions detected by our method quantitatively with the ground truth, which was manually detected hand regions, we calculated the recall and precision rates as follows. Let Φ be the hand regions extracted with our method and Ψ be the

Table 1: Cooking activities in the experiments.

Coarse categories	Fine-grained categories
Cutting	Slicing, Cutting into long thin stripes, Cutting into small pieces Cutting coarsely, Cutting into thin rectangles
Mixing	Kneading by hand, Mixing with whipper, Mixing with chopsticks
Peeling	Peeling with peeler, Peeling with kitchen knife
Frying	-
Washing	-

Table 2: Quantitative results of hand region detection. In this table, $\oplus$ indicates the concatenation of features and we show average F-measure values, which are calculated from the recall and precision rates of hand region detection.

Combination of features	F-measure
$\mathbf{C}_{\mathrm{RGB}} \oplus \mathbf{C}_{\mathrm{HS}} \oplus \mathbf{C}_{\mathrm{AB}}$	0.956
$\mathbf{C}_{\mathrm{RGB}} \oplus \mathbf{C}_{\mathrm{HS}} \oplus \mathbf{C}_{\mathrm{AB}} \oplus \mathbf{T}_{\mathrm{Gabor}}$	0.957
$\mathbf{C}_{\mathrm{RGB}} \oplus \mathbf{C}_{\mathrm{HS}} \oplus \mathbf{C}_{\mathrm{AB}} \oplus \mathbf{T}_{\mathrm{Gabor}} \oplus \mathbf{L}_{\mathrm{block}}$	**0.962**

manually detected regions. Recall rates $R(\Phi, \Psi)$ and precision rates $P(\Phi, \Psi)$ of hand region detection and F-measure value $F(\Phi, \Psi)$ were calculated as follows.

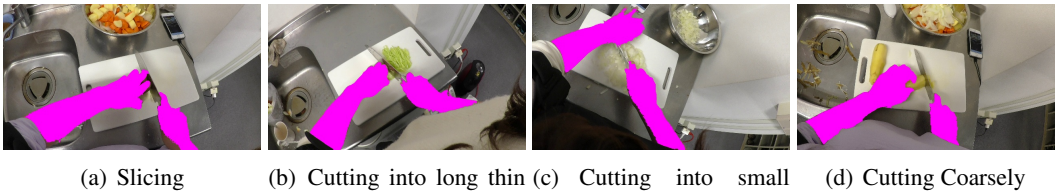
$$R(\Phi,\Psi) = \frac{\Phi \cap \Psi}{\Psi},\ P(\Phi,\Psi) = \frac{\Phi \cap \Psi}{\Phi},\ F(\Phi,\Psi) = \frac{2 \cdot R(\Phi,\Psi) \cdot P(\Phi,\Psi)}{R(\Phi,\Psi) + P(\Phi,\Psi)} \tag{1}$$

Table 2 shows the quantitative results of hand region detection with F-measure values. The combination of color, texture, and location information allowed us to detect the hand region with the highest F-measure. Figure 5 shows the qualitative results. Some background objects were erroneously detected as hand region; for example, hand regions reflected in the basin were erroneously detected because they appeared similar to real hand regions. Therefore, in the future work, we need to improve hand region detection accuracy by using features that are more discriminative or recent advanced deep learning approaches [25, 26]. These results confirm that our method enables us to distinguish hand regions from manipulated objects with qualitatively high performance by combining color, texture, and location information.

In addition, we compare our method with recent approaches [25, 26]. Advances in deep learning have meant that many deep learning hand detection methods have been proposed and have achieved high detection performance [25, 26]. However, a huge amount of data is needed to achieve this performance. In contrast, our approach is simple, which allows us to detect hand region with a small number of videos for training. However, in this experiment, we evaluated our method with small number of test samples in a restricted situation. Therefore, in future work, we need to evaluate our method with large number of test samples, such as EPIC Kitchen dataset [27].

4.2 Experiments of Activities Recognition

Next, we evaluated the recognition performance of our method with 12 cooking activities. We employed three subjects for training and one subject for testing; that is, we utilized 36 videos for training and 12 videos for testing.



(a) Slicing (b) Cutting into long thin stripes (c) Cutting into small pieces (d) Cutting Coarsely

(e) Cutting into thin rectangles (f) Kneading by hand (g) Mixing with whipper (h) Mixing with chopsticks

(i) Peeling with peeler (j) Peeling with kitchen knife (k) Frying (l) Washing

Figure 5: Example of qualitative results of hand region detection in 12 cooking activities.

Table 3: Experimental results for recognition of 12 cooking activities.

Test Subject	Accuracy [%]
subject A	38.9
subject B	33.0
subject C	45.2
subject D	38.9

We conducted the evaluation by chang-ing the testing subject. In addition, we utilized all features (color, texture, and location information) for hand detection and we extracted hand motion features from 20 successive frame images. Therefore, we collected 13440 features (= 3360 features × 4 subjects) and we utilized $\{500, 50, 10, 12\}$ FCMLNN neurons.

Table 3 shows the experimental results. Our method achieved an average recognition rate of 39.0% for 12 cooking activities, which indicates that we need to improve the recognition performance with features that are more discriminative. Figure 6 shows the confusion matrix for the recognition of 12 activities with test subject C, who achieved the highest recognition performance. The “cutting” activities were difficult to distinguish and “frying” activities were erroneously recognized as “Mixing with chopsticks” because their motions are similar. Our method focuses on only repetitive motions for recognition. Therefore, to improve the recognition accuracy of motions like cutting, we also need to recog-nize the shape of ingredients being cut. In addition, similar to the detection experiment, due to advances in deep learning, many approaches for activity recognition have been proposed [28, 29, 30]. Although these methods still have the problem of requiring huge numbers of data samples, they may improve the recognition performance. Therefore, in the future work, we intend to import such approaches to

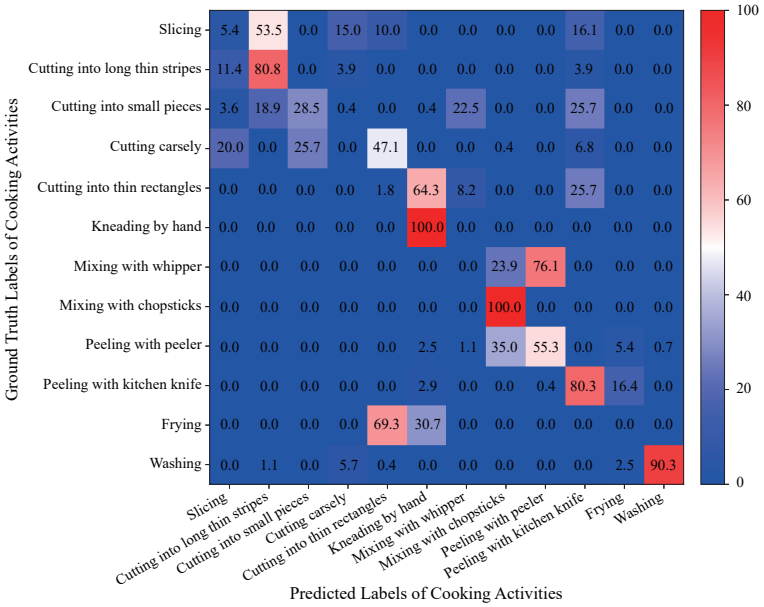


Figure 6: Confusion matrix for recognition results for 12 cooking activities with test subject C who had the highest recognition accuracy.

recognize the fine-grained cooking activities precisely. our method accurately distinguished activities with characteristic motion, such as “mixing with chopsticks”, “kneading by hand”, and “washing.” In addition, although our method still recognizes fine-grained cooking activities insufficiently, it can classify the cooking activities in to coarse categories, such as cutting, peeling, and mixing, with an average recognition rate of 59.7%. These results suggest that our proposed can represent hand motion efficiently and has the potential to classify the cooking activities with characteristic motions.

5 Conclusions

We proposed a novel method for detecting hand regions in egocentric videos based on color, texture, and location information. The experimental results confirmed empirically that our method achieved qualitatively and quantitatively high detection performance. In addition, we presented a simple, efficient hand motion representation method base on a thinning algorithm and a cooking activities recognition method based on a FCMLNN. The experimental results for recognition of 12 fine-grained cooking activities in five coarse categories performed four subjects showed that our FCMLNN achieved a maximum accuracy of 45.2%.

In future work, we intend to improve the performance of hand region detection and cooking activities recognition. We will also investigate cooking activities recognition with more activities in egocentric videos, taken in various illumination environments, and with more subjects.

References

[1] A. Fathi, X. Ren, and James M. Rehg. Learning to Recognize Objects in Egocentric Activities. In *Proc. of CVPR*, 2011.

[2] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding Egocentric Activities. In *Proc. of ICCV*, 2011.

[3] K. Inoue, M. Ono, and M. Yoshioka. Hand Detection and Cooking Activities Recognition in Egocentric Videos. In *Proc. of ACIS*, 2016.

[4] P. Kakumanu, S. Makrogiannis, and N. Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122, 2007.

[5] E. B. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky. Visual Hand Tracking Using Nonparametric Belief Propagation. In *Proc. of CVPRW*, 2004.

[6] I. Oikonomidis, N. Kyriazis, and A. Argyros. Markerless and Efficient 26-DOF Hand Pose Recovery. In *Proc. of ACCV*, 2010.

[7] C. Li and K. M. Kitani. Pixel-level Hand Detection in Ego-Centric Videos. In *Proc. of CVPR*, 2013.

[8] C. Li and K. M. Kitani. Model Recommendation with Virtual Probes for Egocentric Hand Detection. In *Proc. of ICCV*, 2013.

[9] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, 2004.

[10] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Retrieving similar styles to parse clothing. *IEEE TPAMI*, 37(5):1028–1040, 2015.

[11] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *Proc. of ICCV*, 2014.

[12] Y. Li, A. Fathi, and J. M. Rehg. Learning to Predict Gaze in Egcentric Video. In *Proc. of ICCV*, 2013.

[13] Y. Li, Z. Ye, and J. M. Rehg. Delving into Egocentric Actions. In *Proc. of CVPR*, 2015.

[14] H. Wang, A. Kläser, C. Schmid, and C. L. Liu. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *IJCV*, 103(1):60–79, 2013.

[15] S. Escalera, X. Baró, J. González, M. A. Bautista, M. Madadi, M. Reyes, V. P. López, H. J. Escalante, J. Shotton, and I. Guyon. ChaLearn Looking at People Challenge 2014: Dataset and Results. In *Proc. of ECCVW*, 2014.

[16] K. Inoue, T. Shiraishi, R. Matsuoka, and M. Yoshioka. Investigation of Japanese Dyanamic Finger-spelled Sign Language Recognition with RGB-D camera. In *Proc. of FCV*, 2016.

[17] B. Soran, A. Farhadi, and L. Shapiro. Generating Notifications for Missing Actions: Don't forget to turn the lights off! In *Proc. of ICCV*, 2015.

[18] Y. Zhou and T. L. Berg. Temporal Perception and Prediction in Ego-Centric Video. In *Proc. of ICCV*, 2015.

[19] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE TASSP*, 26(1):43–49, 1978.

[20] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching. In *Proc. of CVPR*, 2015.

[21] Z. Li and J. Chen. Superpixel Segmentation using Linear Spectral Clustering. In *Proc. of CVPR*, 2015.

[22] N. Otsu and T. Kurita. A new scheme for practical flexible and intelligent vision systems. In *Proc. of CV*, pages 431–435, 1988.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. of NIPS2012*, 2012.

[24] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio. Maxout Networks. In *Proc. of ICML*, 2013.

[25] A. U. Khan and A. Borji. Analysis of Hand Segmentation in the Wild. In *Proc. of CVPR*, 2018.

[26] K. Roy, A. Mohanty, and R. R. Sahay. Deep Learning Based Hand Detection in Cluttered Environment Using Skin Segmentation. In *Proc. of ICCVW*, 2017.

[27] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *Proc. of ECCV*, 2018.

[28] H. Kwon, Y. Kim, J. S. Lee, and M. Cho. First Person Action Recognition via Two-stream ConvNet with Long-term Fusion Pooling. *Pattern Recognition Letters*, 112, 2018.

[29] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal Multiplier Networks for Video Action Recognition. In *Proc. of CVPR*, 2017.

[30] S. Urabe, K. Inoue, and M. Yoshioka. Cooking Activities Recognition in Egocentric Videos Using Combining 2DCNN and 3DCNN. In *Proc.of CEA/MaDiMa*, 2018.