Comparative Evaluation of Input Features Used for Deep Neural Networks to Recognize Semantic Indoor Scene from Time-Series Images Obtained Using Mobile Robot

Hirokazu Madokoro, Hanwool Woo, and Kazuhito Sato

Department of Intelligent Mechatronics
Faculty of Systems Science and Technology
Akita Prefectural University
Yurihonjo City, Akita, Japan
Email: madokoro@akita-pu.ac.jp

***Abstract*—Human living indoor environments are changing continuously according to our various lifestyles and activities. Human-symbiotic robots require advanced capabilities of environmental understanding and adaptation. For robotic environmental adaptation, numerous machine-learning-based approaches have been proposed. Moreover, numerous types of features such as brightness, edges, texture, etc. have been used for learning networks. This study was conducted to evaluate combinations of supervised-learning-based indoor scene recognition methods and their input features. This paper presents a framework to provide image features of three types according to learning strategies. The experimentally obtained results evaluate using two open benchmark datasets revealed suitable combinations of input features including weights obtained from category maps of Counter Propagation Networks (CNNs) used for Deep Neural Networks (DNNs). We demonstrate a suitable combination of features from scene images used for semantic indoor scene recognition. Particularly, higher recognition accuracy is obtainable using original time-series images for learning with CNNs.**

Keywords–bags of features; category maps; convolutional neural networks; counter propagation networks; self-organizing maps; and semantic indoor scene recognition.

I. INTRODUCTION

Human vision has a gazing mechanism that selects attention-gathering information from a huge amount of information: around 10^9 bit/s [1]. Treisman et al. defined visual saliency as a bottom-up target extracting mechanism based on physiological knowledge and perception with visual feature attention [2]. Koch et al. [3] proposed Saliency Maps (SMs) as a conceptual model of visual saliency. Subsequently, Itti et al. [4] implemented SMs as a computational model for computer-aided processing of images. Applications using saliency models have been proposed widely for computer vision, machine vision, robot vision, collision detection, autopilot, visual perception, and various recognition systems [5]. Using salient objects as visual landmarks in an environment is regarded as highly useful for semantic category recognition used as components that characterize a complex scene [6].

Human living indoor environments are changing continuously according to our various lifestyles and activities. Human-symbiotic robots must have advanced capabilities of environmental understanding and adaptation. Numerous machine-learning-based approaches have been proposed for the adaptation of robots in general environments [7][8][9]. In our earlier study, we proposed a supervised-learning-based scene recognition method using category maps [10]. Our experimentally obtained results revealed that category maps, which visualize relations among scene features, are beneficial for semantic scene recognition.

In conventional machine-learning-based methods, suitable combinational features are extracted in advance. Subsequently, the number of feature dimensions is set as equal using Bag-of-Features (BoF) representation methods. Recently, Deep Neural Networks (DNNs) of various models are fascinating because of their advanced classification and recognition accomplishments [11]. We contemplate that improved accuracy is obtainable for robotic scene recognition using DNNs. However, no scene recognition result has been reported for DNNs trained using weights extracted from category maps.

Saliency-based features are used widely for outdoor and indoor scene classification and for recognition tasks. In one earlier study of saliency-based object recognition, Shokoufandeha et al. [12] examined an SM Graph (SMG) that extracts object saliency regions in several scales using wavelet transformation. Walther et al. [13] produced a biologically plausible model based on SMs for detecting objects from natural scenes. They used a Scale-Invariant Feature Transform (SIFT) [14] descriptor for extracting and describing object features. For outdoor scene recognition, Agrawal et al. [15] described a method to specify accurate positions using cost effective sensors simultaneously combined with GPS. As a challenging reason for indoor scene recognition, Quattoni et al. [8] demonstrated that vast indoor scenes are characterized by objects. It is limited to scenes that are characterized by spatial properties.

Fornoni et al. [16] explained an image classification method based on saliency used for indoor semantic scene recognition. For their method, they used SIFT and Support Vector Machines (SVMs) [17] as a feature descriptor and as a classifier. Botterill et al. [18] proposed a real-time detection method of similar scenes for position estimation used for a mobile robot. They used low-dimensional codebooks combined with a rapid descriptor based on Speeded-Up Robust Features (SURF) [19]. Their method achieved not only rapid object extraction and recognition, but also position estimation in real time for 30 fps, which is an ordinary video frame rate.

For one earlier study using feature descriptors and DNNs, Sachdeva et al. [20] compared the respective accuracies of

 ISBN: 978-1-61208-752-8

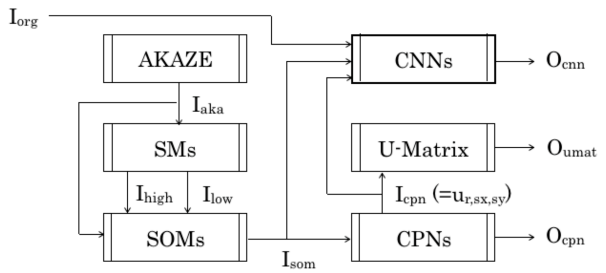

Figure 1. Whole system structure of our proposed framework including data flows among several algorithms.

their proposed model using SIFT and of Convolution Neural Networks (CNNs). They reported that CNNs, which learned using original images, achieved superior accuracy to those which learned using SIFT features with BoF. Mundher et al. [21] proposed a facial expression recognition method using fully connected CNNs combined with dense SIFT. The accuracy of their method was superior to that achieved using the conventional method using CNNs trained using original image features. Both results demonstrated a different tendency for selecting features from scene images used for learning data of CNNs. Actually, the main advances in semantic indoor scene relies on making use of state-of-the-art DNNs. However, we consider that the limitation of DNN-based approaches is inarticulate between input image features and recognition accuracies.

This study was conducted to evaluate combinations of machine-learning-based semantic indoor scene recognition methods and their input features. This paper presents a framework for providing image features of three types according to learning strategies. The experimentally obtained results evaluated using two open benchmark datasets revealed suitable combinations of input features including weights obtained from category maps used for DNNs. We demonstrate a suitable combination of features from scene images used for learning data of CNNs.

The rest of the paper is structured as follows. Sections II and III present our proposed method and an experimental setup including benchmark datasets and evaluation criteria, respectively. Subsequently, Section IV presets evaluation experimental results with discussion. Finally, Section V concludes and highlights future work.

II. Proposed Method

A. Whole architecture

Our proposed supervised-learning-based semantic indoor scene recognition method comprises the following six steps:

1) description of Accelerated KAZE (AKAZE) features,
2) selection of salient regions using SM,
3) generation of BoF using Self-Organizing Maps (SOMs),
4) creation of category maps using Counter Propagation Networks (CPNs),
5) extraction of category boundaries using U-Matrix,
6) and recognition of semantic scenes using DNNs.

Figure 1 depicts the whole system architecture of our proposed framework including data flows among the respective algorithms used for the system. First, local features are extracted from an original input image I_{org} using AKAZE [23] for feature description. Subsequently, high-saliency or low-saliency regions are divided using SMs. Herein, I_{aka} and I_{sm} respectively denote AKAZE features and an image mask of SMs. AKAZE features on high saliency regions I_{high} and those on low saliency regions I_{low} are defined as

$$I_{high} = I_{org} \wedge I_{aka} \wedge I_{sm}, \quad (1)$$

$$I_{low} = I_{org} \wedge I_{aka} \wedge \overline{I_{sm}}. \quad (2)$$

Our method adopts SOMs [24] for BoF. Subsequently, codebooks are created from I_{aka}, I_{high}, and I_{low}. Letting I_{som} a be histogram of SOMs as codebooks, then using I_{som} as input features, category maps are created with CPNs [25]. Letting I_{cpn} be weights of CPNs, then category boundaries are extracted from I_{cpn} using a U-Matrix. For the comparison of recognition accuracies, I_{org}, I_{som}, or I_{cpn} are used as input features for CNNs.

B. AKAZE descriptor

For conventional generic object recognition, SIFT [14] has been used widely for use as an outstanding descriptor of local features. Actually, SIFT descriptors are robust for rotation, scale, position, and brightness changes not only from a static image, but also from dynamic images. Alcantarilla et al. [22] proposed KAZE using nonlinear scale space as a feature that exceeded the SIFT performance. Moreover, they proposed Accelerated-KAZE (AKAZE) [23], which accelerated construction of nonlinear scale spaces of KAZE. In contrast to SIFT, AKAZE was demonstrated as being approximately three times faster, although maintaining equivalent performance and accuracy. Therefore, we use AKAZE, which is suitable for indoor environments where environmental changes occurred frequently.

C. SMs

Briefly, the procedures of SMs include the following five steps. First, a pyramid image is created from I_{org}. Second a Gaussian filter is applied to the pyramid image. Third, images of the respective components of color phase, brightness, and direction are created. Fourth, Feature Maps (FMs) are created as visual features of each component with center-surround and normalization operations. Finally, SMs are obtained from a Winner-Take-All (WTA) competition for the linear summation of FMs.

D. BoF

For this study, we used SOMs to create codebooks. Fig. 2 presents our codebook creation procedure from I_{aka} as BoF. The following is the SOM learning algorithm.

Let $x_p(t)$ be output from the input layer unit p $(1 \leq p \leq P)$ at time t. As input features, I_{aka}, I_{high}, and I_{low} are given to $x_p(t)$. Let $w_{p,q}(t)$ be a weight from p to mapping layer unit q $(1 \leq q \leq Q)$ at time t. Herein, P and Q respectively denote the total numbers of input layer units and mapping layer units. Before learning, $w_{p,q}(t)$ are randomly initialized. Using the Euclidean distance between $x_p(t)$ and $w_{p,q}(t)$, a winner unit $c_q(t)$ is sought for the following.

$$c_q(t) = \underset{1 \leq q \leq Q}{\operatorname{argmin}} \sqrt{\sum_{p=1}^{P} (x_p(t) - w_{p,q}(t))^2}. \quad (3)$$

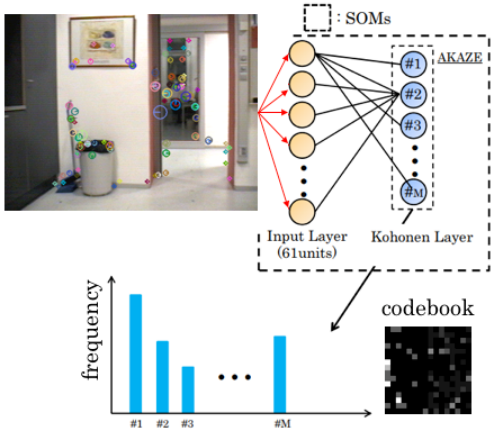

Figure 2. Codebook creation procedure using SOMs from I_{aka} as BoF.

A neighborhood region $\psi_{som}(t)$ is set from the center of c_q as the following.

$$\psi_{som}(t) = \lfloor \psi_{som}(0) \cdot Q \cdot \left(1 - \frac{t}{Z_{som}}\right) + 0.5 \rfloor, \quad (4)$$

where Z_{som} is the maximum learning iteration. Subsequently, $w_{p,q}(t)$ in $\psi_{som}(t)$ is updated as

$$w_{p,q}(t+1) = w_{p,q}(t) + \alpha(t)(x_p(t) - w_{p,q}(t)), \quad (5)$$

where $\alpha(t)$ is a learning coefficient that is decreasing according to the learning progress.

After learning, test data are entered to the input layer. A winner unit is used for voting to create a histogram as a codebook: I_{som}. We obtained I_{som} of two types: a 1-Dimensional (1D) codebook using a 1D category map and a 2D codebook using a 2D category map. For creating I_{som}, the index of the mapping layer is changed to qx and qy.

E. CPNs

We create a category map using CPNs. For learning CPNs, I_{som} are entered to the input layer of CPNs as input features. Let $y_r(t)$ be output from the input layer unit r $(1 \leq r \leq R)$ at time t. Let $w_{r,s}(t)$ be a weight from r to Kohonen layer unit s $(1 \leq s \leq S)$ at time t. Herein, R and Q respectively denote the total numbers of input layer units and Kohonen layer units. Before learning, $w_{r,s}(t)$ are initialized randomly. Using the Euclidean distance between $y_r(t)$ and $w_{r,s}(t)$, a winner unit $c_s(t)$ is sought for the following.

$$c_s(t) = \underset{1 \leq s \leq S}{\operatorname{argmin}} \sqrt{\sum_{r=1}^{R} (y_r(t) - u_{r,s}(t))^2}. \quad (6)$$

A neighborhood region $\psi_{cpn}(t)$ is set from the center of c_s as the following.

$$\psi_{cpn}(t) = \lfloor \psi_{cpn}(0) \cdot S \cdot \left(1 - \frac{t}{Z_{cpn}}\right) + 0.5 \rfloor, \quad (7)$$

where Z_{cpn} stands for the maximum learning iteration. Subsequently, $u_{r,s}$ and $v_{s,k}$ in $\psi_{cpn}(t)$ is updated as shown below.

$$u_{r,s}(t+1) = u_{r,s}(t) + \beta(t)(y_r(t) - u_{n,m}(t)), \quad (8)$$

Figure 3. Brightness changes in daytime (upper) and nighttime (lower) with similar positions.

$$v_{s,k}(t+1) = v_{s,k}(t) + \gamma(t)(z_l(t) - v^j_{n,m}(t)), \quad (9)$$

where $\beta(t)$ and $\gamma(t)$ are learning coefficients that decrease along with learning progress.

As a learning result, $u_{r,s}$ is used for the input to CNNs. We defined this interface as I_{cpn}.

F. U-Matrix

For this study, we used a 2D Kohonen layer. The unit index s is extended to sx and sy. Category boundaries are extracted from $u_{r,sx,sy}$ using U-Matrix. Based on metric distances between weights, U-Matrix visualizes the spatial distribution of categories from similarity of neighbor units [26]. On a 2D category map of square grids, a unit has eight neighbor units except for boundary units. Assuming U as the similarity calculated using a U-Matrix, then for the component of the horizontal and vertical directions, $U_{h\pm}$ and $U_{v\pm}$ are defined as shown below.

$$U_{h\pm} = \sqrt{\sum_{r=1}^{R} (u_{r,sx,sy} - u_{r,sx\pm1,sy})^2}, \quad (10)$$

$$U_{v\pm} = \sqrt{\sum_{r=1}^{R} (u_{r,sx,sy} - u_{r,sx,sy\pm1})^2}. \quad (11)$$

For the components of the diagonal directions, $U_{d\pm}$ are defined as the following.

$$U_{d\pm} = \frac{1}{2}\sqrt{\sum_{r=1}^{R} (u_{r,sx,sy\pm1} - u_{r,sx\pm1,sy})^2} \quad (12)$$

$$+ \frac{1}{2}\sqrt{\sum_{r=1}^{R} (u_{r,sx\pm1,sy} - u_{r,sx,sy\pm1})^2} \quad (13)$$

G. DNNs

Numerous DNN frameworks are provided. For this study, we used VGG-16 [28] that composed 13 convolutional layers and three fully connected layers. As a mechanism to reduce errors, VGG-16 includes a batch normalization algorithm in each convolutional layer [27]. For general object position identification and classification, VGG-16 demonstrated superior results in large-scale image competitions [28].

 ISBN: 978-1-61208-752-8

III. Experimental Setup

A. Benchmark datasets

Quattoni et al. presented large-scale indoor scene recognition datasets [8]. The database includes 67 indoor categories that collectively include 15,620 images. Although the numbers of images vary among the categories, each category has at least 100 images. Images were obtained using a monocular camera. Therefore, the datasets comprise dispersed images. For a real-world robotics application, it remains a challenging task for a mobile robot to move 67 different places [29].

For this study, we used two open benchmark datasets that comprised time-series images obtained using mobile robots. The first dataset is KTH-IDOL2 [30], which comprises time-series images used for indoor robotics navigation and vision-based position estimation. Indoor scenes are of five categories: a printer area (PA), a corridor (CR), a one-person office (EO), a kitchen (KT), and a two-person office (BO). The image resolution is 320 $\times$ 240 pixels. This dataset includes some object changes because images were obtained at time intervals of up to six months in the same environment. Moreover, rotated images were obtained at PA and KT for providing diverse visual information.

The second dataset includes place recognition datasets [29] that comprise time-series images obtained using a monocular camera on a mobile robot. This dataset includes 17 scene categories: 11 categories at York University and the remaining 6 categories at the Coast Capri Hotel. For this study, the York University sub-dataset and the Coast Capri Hotel sub-dataset are abbreviated respectively as YUSD and CHSD. The respective resolutions of the images are 640 $\times$ 480 pixels.

As common features of both datasets, two robots of different heights were used for image data acquisition. We used images obtained using a higher robot. Both datasets include diverse image appearances with positional shifts because the robot moved a previously setting route with manual operation. Moreover, images are obtained in daytime and nighttime. Fig. 3 shows brightness changes with similar positions. For this study, we used the both illumination condition images.

B. Evaluation criteria

As evaluation criteria, recognition accuracy R is defined as

$$R = \frac{S_{test}}{N_{test}} \times 100, \tag{14}$$

where N_{test} and S_{test} respectively denote the numbers of test images and of correct recognition images. For this study, we used Leave-One-Out Cross-Validation (LOOCV) [31] to evaluate the capability of generalization.

IV. Evaluation Experiment using CPNs

A. Saliency for recognition

We evaluated the relations between recognition accuracy and input features of three types: I_{aka}, I_{high}, and I_{low}. Fig. 4 depicts results obtained from a comparison of the recognition accuracy for KTH-IDOL2. The mean recognition accuracies of I_{aka}, I_{high}, and I_{low} were, respectively, 67.8%, 59.3%, and 61.2%. The recognition accuracy of I_{aka} was 8.5 percentage points higher than that of I_{high} and 6.6 percentage points higher than that of I_{low}. This result revealed that I_{aka} was the highest among three feature patterns.

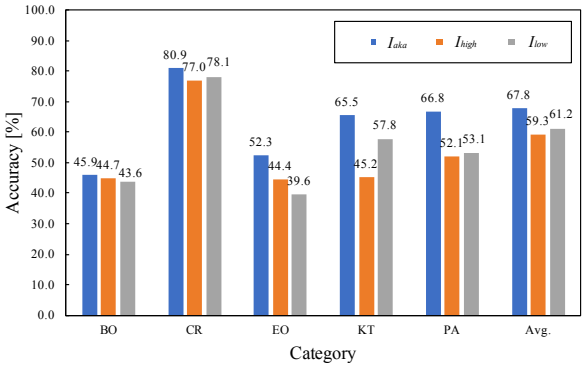

Figure 4. Recognition accuracy in each category for KTH-IDOL2.

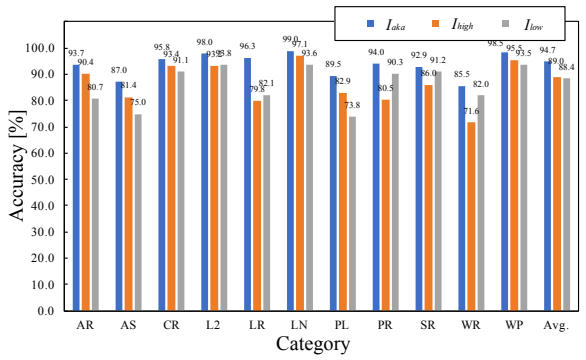

Figure 5. Recognition accuracy in each category for YUSD.

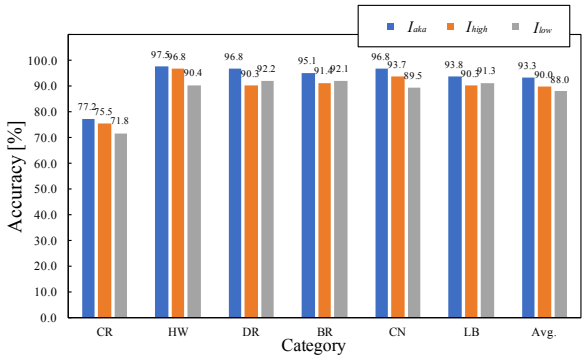

Figure 6. Recognition accuracy in each category for CHSD.

Fig. 5 presents results obtained from a comparison of the recognition accuracy for YUSD. The mean recognition accuracies of I_{aka}, I_{high}, and I_{low} were 94.7%, 89.0%, and 88.4%. The recognition accuracy of I_{aka} was 5.7 percentage points higher than that of I_{high} and 6.3 percentage points higher than that of I_{low}. This result revealed I_{aka} as the highest among three feature patterns.

Fig. 6 presents results obtained for comparison recognition accuracy for CHSD. The mean recognition accuracies of I_{aka}, I_{high}, and I_{low} were 93.3%, 90.0%, and 88.0%. The recognition accuracy of I_{aka} was 3.3 percentage points higher than that of I_{high} and 5.5 percentage points higher than that of I_{low}. Results demonstrated that I_{aka} was the highest among three feature patterns. We obtained the similar tendency for input features. Results also demonstrated that saliency-based

 ISBN: 978-1-61208-752-8

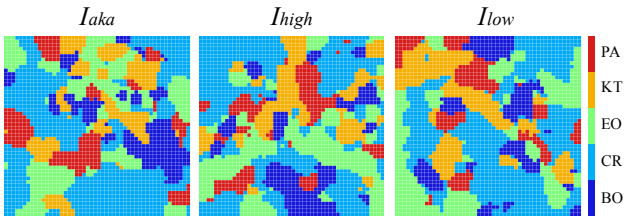

Figure 7. Results of category maps for KTH-IDOL2.

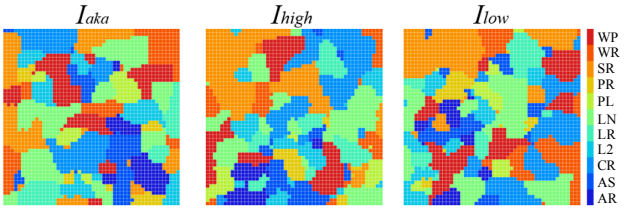

Figure 8. Results of category maps for YUSD.

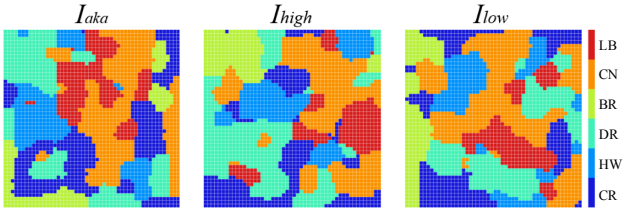

Figure 9. Results of category maps for CHSD.

features dropped recognition accuracy.

B. Category maps

Fig. 7 portrays category maps created from KTH-IDOL2. Unit color patterns correspond to scene category labels. For all feature patterns, scene categories were divided into several independent clusters. Clusters of various shapes and sizes were mixed on the category maps. Moreover, independent clusters consisting of a single unit exist in the maps. Particularly, PA, CR, EO, KT, and CR of I_{aka} respectively comprised 9, 12, 14, 16, and 15 clusters. The CR clusters are larger than those of other categories.

Fig. 8 portrays category maps created from YUSD. As an overall tendency, similar categories are divided into independent clusters that depict scene diversity. Comparison with the result of KTH-IDOL2 reveals that clusters are gathered to particular locations, with few independent units. Fig. 9 displays some category maps created from CHSD. Although positions and sizes differed among categories, the cluster distribution tendency was similar to those of results presented in Fig. 8. The experimentally obtained results revealed that category maps with numerous clusters reflected the complexity of indoor scenes.

C. Category boundary extraction

For analyzing the category relation, we extracted category boundaries using U-Matrix, which calculated the similarity of neighbor units based on the distance of weights between category map units. For enhancing category boundaries, we used the representative automatic image thresholding method

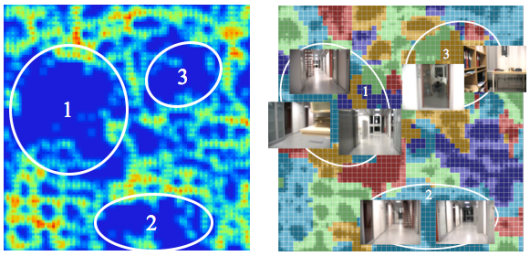

Figure 10. Boundary extraction results and category representative images obtained using U-Matrix.

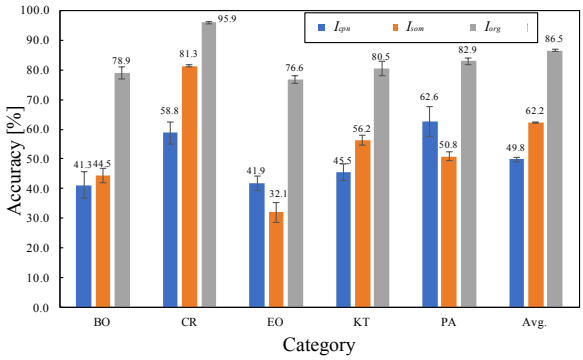

Figure 11. Recognition accuracies and comparison results obtained for each input feature.

reported by Otsu [33]. U-Matrix boundaries are depicted using temperature colors, with high temperature portrayed as red according to the distances among weights.

The left panel of Fig. 10 depicts boundary extraction results for the results depicted in Fig. 7(a). The category map included three independent regions and several slight regions. For the three independent regions, we assigned labels as Boundaries 1–3 according to the order of sizes. The right panel of Fig. 10 portrays a category map with superimposed boundary extraction results and category representative images. Numerous units were labeled to CR in Boundary 1, especially gathered PA images. In Boundary 3, labels were mixed with all categories.

D. Evaluation Experiment using CNNs

For this evaluation experiment, we used KTH-IDOL2 alone because we obtained sufficient recognition accuracies in place recognition datasets using CPNs.

For learning and validation of CNNs, we used input image features of three types: I_{cpn}, I_{som}, and I_{aka}. Fig. 11 presents results obtained from comparison of the respective scene categories. The mean recognition accuracies of I_{cpn}, I_{som}, and I_{aka} with LOOCV were, respectively, 49.8%, 62.2%, and 86.5%. Comparison of the three results indicates the following I_{aka} obtained the highest recognition accuracies in all categories. Regarding details of respective categories, recognition accuracies of I_{som} in BO, CR, and KT were 3.2, 22.5, and 10.7 percentage points higher than those of I_{cpn}. By contract, recognition accuracies of I_{cpn} in EO and PA were, respectively, 9.8 and 11.8 percentage points higher than those of I_{som}.

 ISBN: 978-1-61208-752-8

V. CONCLUSION

For semantic indoor scene comprehension when used for a mobile robot, we evaluated combinations of supervised machine-learning-based methods and input features using AKAZE, SMs, SOMs, CPNs, U-Matrix, and DNNs. After optimizing the parameters and input features, we conducted two experiments using CPNs and CNNs as a recognizer using open benchmark datasets comprising time-series images The experimentally obtained results obtained using CPNs revealed that the mean recognition accuracy of I_{aka} was higher than those of I_{high} and I_{low} in all categories. Several clusters were created on category maps with designated complexity of scenes. The experimentally obtained results obtained using CNNs revealed that higher recognition accuracy was obtainable using original time-series images for learning.

For our future work, we expect to improve the recognition accuracy to reduce false recognition around scene-switching zones. The relation between processing time and recognition accuracy must be assessed with an assumption of adaptation to the real environment. Moreover, future studies must assess recognition when using cameras with spherical lenses. Furthermore, we will implement our proposed framework to a human-symbiotic robot for the conduct of evaluation experiments in an actual environment.

REFERENCES

[1] K. Koch et al., "How Much the Eye Tells the Brain," *Current Biology*, vol. 16, pp. 1428–1434, 2006.

[2] A.M. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[3] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[4] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[5] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[6] H. Madokoro, K. Sato, and N. Shimoi, "Semantic Indoor Scene and Position Recognition Based on Visual Landmarks Obtained from Visual Saliency without Human Effects," *Robotics*, vol. 8, no. 1, pp. 1–24, 2019.

[7] C. Siagian and L. Itti, "Rapid Biologically Inspired Scene Classification Using Features Shared with Visual Attention," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300–312, 2007.

[8] A. Quattoni and A. Torralba, "Recognizing Indoor Scenes," *Proc. Computer Vision and Pattern Recognition*, pp. 413–420, 2009.

[9] A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin, "Context-Based Vision System for Place and Object Recognition," *Proc. IEEE International Conference Computer Vision*, pp. 1023–1029, 2003.

[10] H. Madokoro, Y. Utsumi, and K. Sato, "Unsupervised Indoor Scene Classification Based on Context for a Mobile Robot (in Japanese)," *Journal of the Robotics Society of Japan*, vol. 31, no. 9, pp. 918–927, 2013.
[**?**] S. Lazebnik, C. Schmid and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *Proc. Computer Vision and Pattern Recognition*, pp. 2169–2178, 2016.

[11] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[12] A. Shokoufandeh, I. Marsic, and S.J. Dickinson, "View-Based Object Recognition Using Saliency Maps," *Image and Vision Computing*, vol. 17, pp. 445–460, 1999.

[13] D. Walthera and C. Koch, "Modeling Attention to Salient Proto-Objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[14] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc. IEEE International Conference Computer Vision*, vol. 2, pp. 1150–1157, 1999.

[15] M. Agrawal and K. Konolige, "Real-time Localization in Outdoor Environments using Stereo Vision and Inexpensive GPS," *Proc. 18th International Conference on Pattern Recognition*, pp. 1063–1068, 2006.

[16] M. Fornoni and B. Caputo, "Indoor Scene Recognition using Task and Saliency-driven Feature Pooling," *Proc. British Machine Vision Conference*, pp. 1–12, 2012.

[17] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[18] T. Botterill, S. Mills, and R. Green, "Speeded-up Bag-of-Words algorithm for robot localisation through scene recognition," *Proc. 23rd International Conference Image and Vision Computing*, pp. 1–6, 2008.

[19] H. Bay, T. Tuytelaars, and L.V. Gool, "Surf: Speeded Up Robust Features," *Proc. European Conference on Computer Vision*, pp. 404–417, 2006.

[20] V. Sachdeva et al., "Performance Evaluation of SIFT and Convolutional Neural Network for Image Retrieval," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 8, no. 12, pp. 518–523, 2017.

[21] M. Al–Shabi, W. P. Cheah, and T. Connie, "Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator," *Proc. International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, pp. 139–149, 2017.

[22] P.F. Alcantarilla, A. Bartoli, and A.J. Davison, "KAZE Features," *European Conference on Computer Vision*, pp. 214–227, 2012.

[23] P.F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces," *British Machine Vision Conference*, pp. 1–12, 2013.

[24] T. Kohonen, "Self-Organized formation of Topologically Correct Feature Maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[25] R. H. Nielsen, "Counterpropagation networks," *Applied Optics*, vol. 26, pp. 4979–4983, 1987.

[26] A. Ultsch, "Clustering with SOM U C," *Proc. Workshop on Self-Organizing Maps*, pp. 75–82, 2005.

[27] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *Proc. the 32nd International Conference on International Conference on Machine Learning*, vol. 37, pp. 448–456, 2015.

[28] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Proc. 3rd IAPR Asian Conference on Pattern Recognition*, pp. 730–734, 2015.

[29] R. Sahdev and J.K. Tsotos, "Indoor Place Recognition System for Localization of Mobile Robots," *Proc. 13th Conference on Computer and Robot Vision*, pp. 53–60, 2016.

[30] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "The KTHIDOL2 Database," *Technical Report CVAP304, KTH Royal Institute of Technology, CVAP/CAS*, 2006.

[31] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proc. Fourteenth International Joint Conference on Artificial Intelligence*, vol. 2, no. 12, pp. 1137–1143, 1995.

[32] H. Madokoro, N. Shimoi, and K. Sato, "Adaptive Category Mapping Networks for All- Mode Topological Feature Learning Used for Mobile Robot Vision," *Proc. 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 678–683, 2014.

[33] N. Otsu, "A Threshold Selection Method From Gray-Level Histograms," *IEEE Trans. System, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[34] H. Madokoro and K. Sato, "Visualizing Support Vectors and Topological Data Mapping for Improved Generalization Capabilities," *Proc. IEEE World Congress on Computational Intelligence*, pp. 4226–4232, 2010.

 ISBN: 978-1-61208-752-8