International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.7 (2015), pp.189-198
http://dx.doi.org/10.14257/ijsip.2015.8.7.18

Bio-inspired Approach for Recognition of Single Object in Natural Images against Complex Background

Zuojin Li Liukui Chen* and Chen Gui

College of Electrical and Information Engineering
Chongqing University of Science and Technology
*cqustclk@163.com

Abstract

Recognition of single object in Natural Images usually shows low accuracy and low comprehension ability due to the complexity of the scenes. In contrast, human vision boasts extraordinary perceptive ability, which motives our study on simulation of human visual information pathways. By mimicking the layering processing mechanism, we designed a computing model and method for image perception with the functional features of biological vision. This model can perceive single object against complex background with the accuracy above 90%, as proven in experiments based on the testing sample from the database "caltech101".

Keywords: *Laying processing; Computing model; Natural image; Single object recognition*

1. Introduction

In recent years, cognitive informatics has become a hot issue in the interdisciplinary field of life science and computer science. Cognitive informatics (CI) is a transdisciplinary enquiry of computer science, information science, cognitive science, and intelligence science that investigates into the internal information processing mechanisms and processes of the brain and natural intelligence, as well as their engineering applications in cognitive computing [1,2]. As we all know, the cognitive ability of computers is incomparable to that of humans and primates. Therefore, how to improve this ability learning from biological vision is a focus and also a great challenge in computer vision [3-7].

The existing researches on biology show that primates possess two vision pathways in information processing, namely "where pathway" and "what pathway" [8-10]. The latter is responsible for recognizing objects, such as faces, items, etc. and its mechanism are important for machine vision to learn from. In the "what pathway", information starts from Area V1, via Areas V2 and V4, and reaches Area IT, where the tasks of information processing and recognition finish [11]. During this process, the analysis of the visual system to the receptive fields is layered and progressively more complex, showing gradually more abstract and integral presentation of information features. It is because of these highly complicated layering features that biological visual perception boasts strong robustness and fast response. This layered information processing mechanism provides a new method for single object perception of machine vision and also motivates this study.

2. Related Work

In recent years, the layered computing modelling inspired from the information processing mechanism of biological visual cortex has drawn tremendous attention, especially in computer vision field. B. Heisele and B. Leung etc [12, 13] proposes a series of layered computing models, achieving face and vehicle detection in natural

ISSN: 2005-4254 IJSIP

scenes, which possess learning ability as biological vision. M. Weber *etc.* [14-17] also posts a couple of layered neural network computing models based on convolution operation, demonstrating high accuracy in figure and face recognition; B. Leibe *etc.* [18-21] establishes a perception model based on contour features and block learning. These models all possess layered information processing function, but not perceptive selectivity and invariant feature extraction function as the "what pathway" of biological vision. Therefore, they are not sufficient in terms of their biological visual mechanism, or, in other words, far incomparable to the robust perception of biological vision [22], by far, is the computing model inspired most in-depth from the biological vision, which includes not only two vision pathways, but also integrates the features at low, middle and high levels. It also possesses memorizing and learning ability as human brain, which enables it to search for objects even in complex environment. The shortcoming, however, is the complexity of algorithm structure and the enormity of the computing workload, which makes its application in engineering almost impossible. D. G. Lowe [23] proposes a computing model with perceptive invariant features, referred to as SIFT features, which is a popular method in computer vision application. Compared with other models of this type, this model shows better SIFT features in detecting learnt samples, but low accuracy to unlearnt samples. Especially, when applied in perceiving multiple objects against natural complex background, a lot of noisy points are hidden in its invariant features and the computing process becomes quite time consuming and shows low robustness.

Aiming at overcoming the above mentioned shortcomings, this paper adopts the information processing mechanism of biological vision and proposes a layering computing method suitable for single object recognition in natural scenes. The following part is structured as below: Section III presents the computing method simulating vision mechanism; Experiments and discusses are given in Section V. Finally, we conclude this paper in Section VI.

3. Computing Method

M. Riesenhuber and T. Serre *etc.* [24, 25] propose a theory that biological cortex adopts layered method in information processing and pattern recognition from the quantitative respective based on psychological experiments data on objects recognition of biological vision. This theory holds that: 1) the layered processing of vision information aims to extract the invariance of objects' positions and scales, and gradually achieve rotational invariance; 2) with the layered processing, the receptive field of cells change from simple to complex and consequently form the stimulation features of complex cells' receptive field; 3) primary recognition of vision is the processing results of feedfoward information; 4) IT cells has plasticity and learning ability, which enables the cortex to respond quickly to classic objects. Based on these conclusions, we propose an object recognition modelling with cortex biological features as shown in Figure 1. This model simulates the layering mechanism of biological vision, composed of 5 levels of different feature functions.

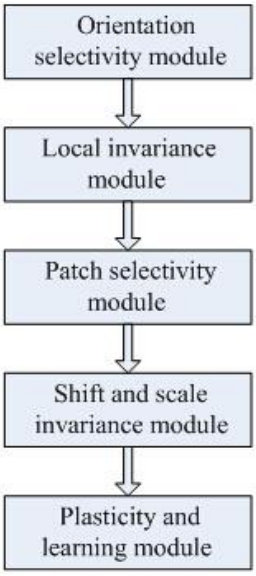

Figure 1. Computation Model of Ventral Stream of Visual Cortex

The computing method based on this model is as follows:

Orientation Selectivity Module: the receptive fields of simple cells in the low-level visual cortex have selectivity to directional sensitivity, which can be mathematically simulated with two-dimensional Gabor function. To simplify the calculation, we use the real part of the Gabor function to express, as shown in the following equation [26]:

$$g(x,y)=\exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\times\cos\left(\frac{2\pi}{\lambda}x'\right), \tag{1}$$

where

$$\begin{cases} x'=x\cos\theta+y\sin\theta \\ y'=-x\sin\theta+y\cos\theta \end{cases}, \tag{2}$$

θ represent the direction of the filter, which can point to any direction by rotating the axis. This is exactly the feature of sensitivity of cells receptive field to directional selectivity. In order to match the biological experiment data and facilitate calculation, we choose θ=0°，45°，90°，135°; γ represents the width-to-height ration of the Gussian function, usually γ=0.3; λ represents the wavelength of the modulation function; σ, the covariance of the Gussian function, decides the effective zone of the filter.

In order to achieve the invariance and selectivity of vision to different scales, we adopts a pyramid scale filter set in this step. The values taken for λ and σ in each scale space are mainly based on data of some biological experiments on vision perception [26]. The scale space starts from 7×7, with the step length of 2 pixels and maximum scale of 37×37.

Local Invariance Module: local invariance can be interpreted as the stimulation responses of complex cells' receptive fields do not change to the external scales. To make the response value constant within a certain scope, this value should be the maximum in the local scope; otherwise, there always exists a bigger value when the external scales change, which is contradictory to the invariance feature of biological vision. Here we decide the local scope as two adjacent scale spaces with an area of $l\times l$, where l is the average value of the scales in the two spaces. The invariance equation is expressed as below:

$$L=\max(g_1,\cdots,g_m), \tag{3}$$

where, g_i is the cell's response value of the orientation selectivity module.

Patch Selectivity Module: in biological visual system, the function of this module is to select objects according to their features stored in the brain, whose accuracy depends on the similarity between the objects and targets. The similarity can be expressed mathematically as the Euclidean distance, with Gaussian kernel-based radial function, as shown in Equation (4):

$$P = \exp\left(-\|\bullet\|^2 / 2\delta^2\right), \tag{4}$$

where δ is modulation parameter of Gaussian function shape, which is generally defined as the average value of the distance between the cluster center and the sample; $\|\bullet\|$ is the Euclidean distance between the input feature and the center of the primary function, where the center of the primary function is the center of the sample. The sample patch is obtained by random selection from the Set L, with the size of 6×6. In order to reduce the feature dimensions, we normally let the sample number $K = 20\sim 60$ in the experiment. The exact number will be given later in the Experiment section. After calculation, P is the column vector of $K\times 8$.

Shift and Scale Invariance Module: With the performance of the previous modules, visual features have been selected locally and their invariance has been extracted. When perceiving objects, the visual system should form an overall feature, or the global invariance. Similar to the principle of expressing local invariance, we should also obtain the maximum value to represent the invariance of the position and scale of pixels. In other words, we should choose K maximum values from P as the invariance feature of the input image.

Plasticity and Learning Module: Plasticity and learn-ability are typical features of biological vision, which can be realized by the artificial optic neural network. To simplify the calculation, we use two-layer neural network to simulate the function column structure of the IT area and adopt BP algorithm for weight learning of neural network, whose structure is shown in Figure 2.

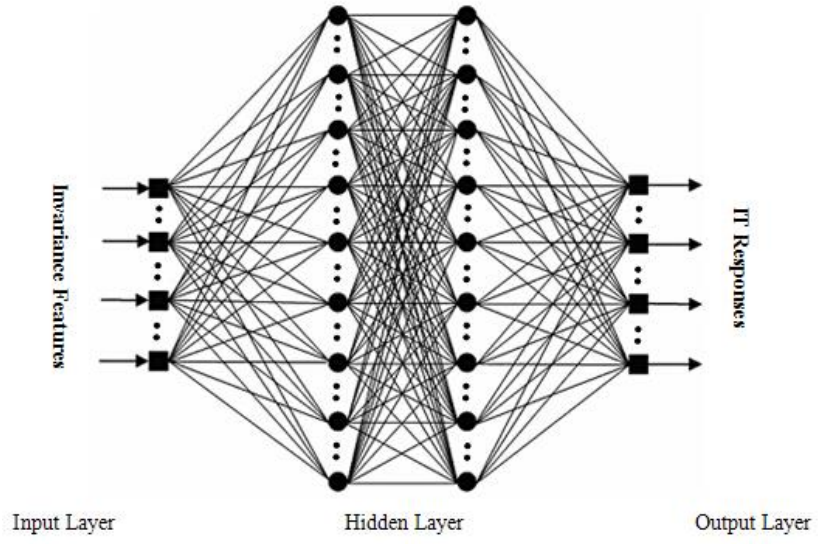

Figure 2. Plastic and Learning Nuero Network Computing Model

The number of nodes on the input layer of the invariance feature is K, while that on the output layer equals the number of object categories or pixels of the input

images. The number of hidden layers is decided by the number of nodes on the input and output layers, which can be expressed in Equation (5).

$$S = \omega \bullet \sqrt{m + n + 0.35 + 12} \quad (5)$$

where S is the number of nodes in the hidden layer and m and n represent the numbers of nodes in the input and output layers respectively. ω is usually valued between5-7.

4. Experiment and Discussion

In order to verify the robust recognition effect of this method under a complex background, we choose to use the common natural scene photo database "caltech101" [16] published by the California Institute of Technology. This database contains photos of 101 common and typical objects in natural environment, such as cars, footballs, cameras, planes, etc., as shown in Fig. 3, there are four sample objects from this database. In this database, there are 40-800 photos for each object, with the image size of 300×200 pixels; for each object, some photos only contain the object itself not the environment, as shown in Figure 3(d); while some contains both the object and the environment, as shown in Figure 3(a), (b) and (c).

(a) (b)

(c) (d)

Figure 3. Object Image Samples from Caltech 101

In the experiment, we reduced the height of every picture to 200 pixels to unify their sizes, and then reduce or increase their width according to their original aspect ratios. When compiling programs, we use "+1" to represent correct recognition result and "-1", incorrect ones. All the result values range from "-1" to "+1" during the program run, and we take "0" as the threshold. As the number of pictures differs from category to category, we conduct training on 5%, 10%, 15%, 20%, 30% and 40% of the sample, with the rest being tastes. Fig. 4 shows the statistical histogram of accurate recognition for each category with different amount of training. In this graph, the horizontal axis represents training percentages and with the same training percentage, we compare the recognition results for the same object with different methods. "A" stands for Learning Components with Support Vector Machines [12]; "B", Component-based [13]; "C", Unsupervised learning of models [14]; "D", unsupervised scale-invariant learning [15]; and "E", the method proposed in this paper. The vertical axis represents the correct recognition ratio (the highest is 1) and the box plots demonstrate the performance of different methods on objects under No.101 category of this database. In this experiment, we only choose 30 dimensional invariant features randomly, as the complete computing is extremely time consuming.

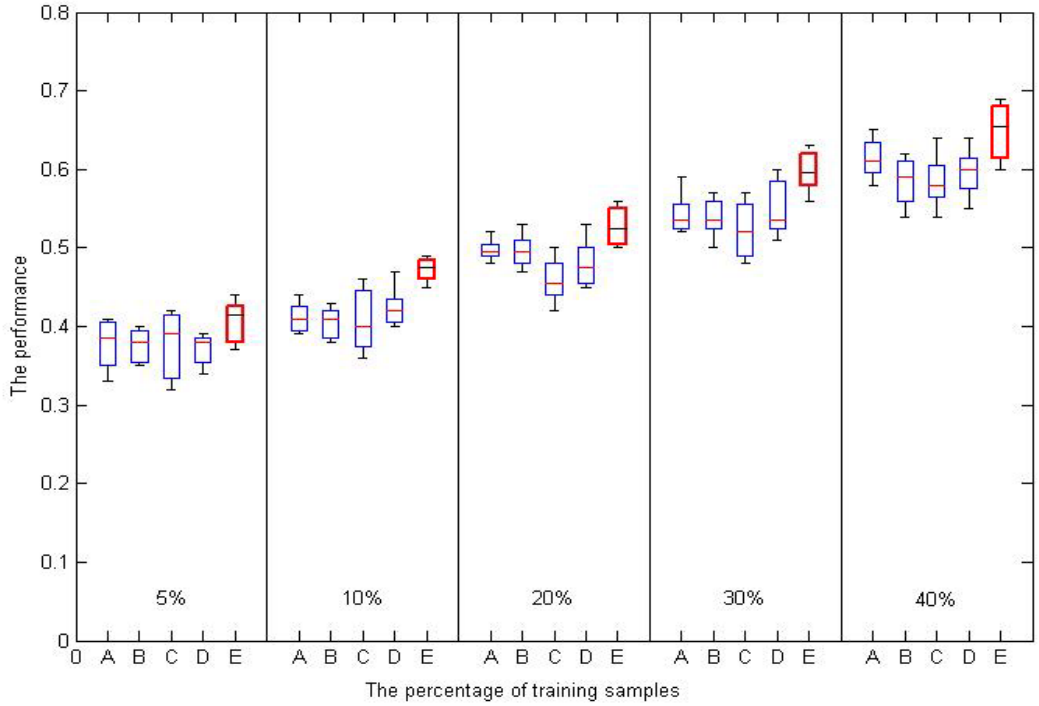

Figure 4. Object Identification Rate of Caltech101 Database with Training Percentage of 5%, 10%, 20%, 30% and 40%

Furthermore, in order to further prove the practicality of this method, we also choose objects under complex background and compare the results with the most popular method, SIFT. As SIFT is a method for invariant feature extraction, we contrast it with the feature layers of this method. Objects chosen for testing include lion faces, "stop" signs, boats, cups and motorcycles, totalling 300 natural pictures, some of them shown in Figure 5., The correct recognition ratios of this method and "SIFT+SVM" [23] method under different feature dimensions are shown in Figure 6. We can see that our method achieves better results than the latter.

Figure 5. Test Samples Object of Five in Caltech101 Database

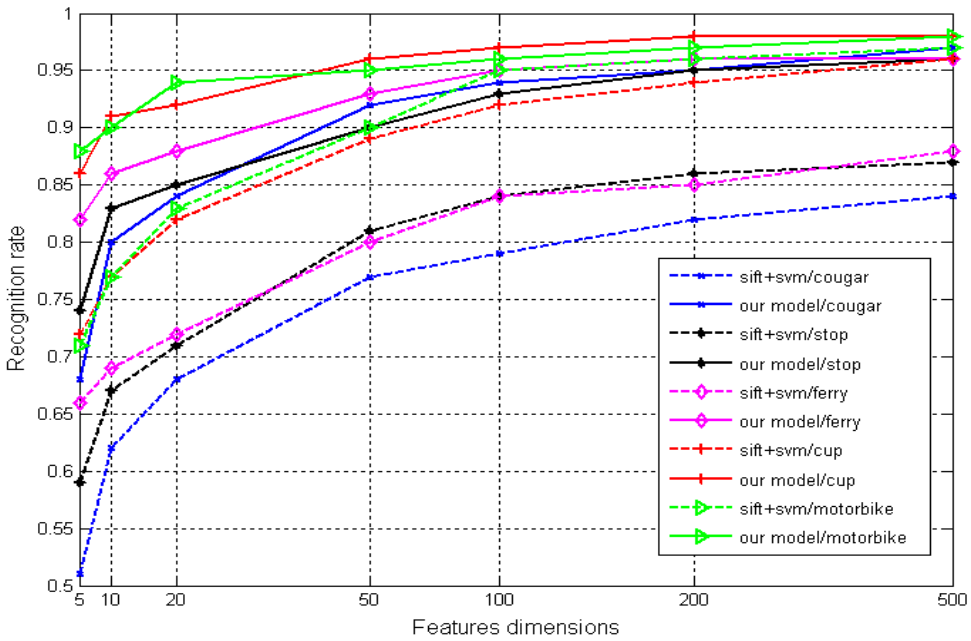

Figure 6. Correct Recognition Ratios of This Method and Other Methods under Different Feature Dimensions

5. Conclusion

The complex background makes the recognition for natural images extremely difficult for machine vision, while human vision is highly efficient in this. This paper learns from human vision and proposes a new method simulating its layering information processing mechanism, which realizes single object identification under complex background. Detailed computing methods are explained and identification testing is conducted on 101 typical objects in "Caltech 101" database. The 90% accuracy proves that our method enjoys robustness in perceiving natural images and shows higher accuracy compared with other methods. It is believed to be able to improve the engineering application of machine vision. Following work will be carried out on the suitability of this method on multi-objects identification in natural scenes.

Acknowledgements

This work is partially supported by Scientific and Technological Research Program of Chongqing (Grant No. KJ131423, No. KJ131422 and KJ132206), Doctor Special Project from Chongqing University of Science and Technology (Grant No. CK2011B05 and CK2011B09) and Natural Science Foundation of Chongqing (Grant No. cstcjjA40041). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] Y. Wang, "Perspectives on the Field of Cognitive Informatics and its Future Development", International Journal of Cognitive Informatics and Natural Intelligence, vol. 5, no. 1, **(2011),** pp. 1–47.
[2] Y. Wang, "On cognitive models of causal inferences and causation networks", International Journal of Software Science and Computational Intelligence, vol. 3, no. 1, **(2011),** pp. 50-60.
[3] Q. Tang, N. Sang, and H. Liu, "Learning to detect contours in natural images via biologically motivated schemes", **(2013)** September, pp. 123-126, ICIP.
[4] H. Li, H. Li, Y. Wei and Y. Tang, "Sparse-based neural response for image classification", Neurocomputing, vol. 144, **(2014),** pp. 198-207.
[5] S. Li, X. Niu, Z. Wang and H. Shi, "A study on image representation method based on biological visual mechanism", **(2012)** November, pp. 1283-1288, ICAMME.
[6] D. Hu, J. Li, S. X. Yang and S. Gregori, "A multiple feature fusion model for image segmentation based on hierarchical visual pathway", Advances in Information Sciences and Service Sciences, vol. 4, no. 8, **(2012),** pp. 274-282**.**
[7] Y. Hatori and S. Ko, "Early representation of shape by onset synchronization of border-ownership-selective cells in the V1-V2 network", Journal of the Optical Society of America A: Optics and Image Science, and Vision, vol. 31, no. 4, **(2014),** pp. 716-729**.**
[8] L. G. Ungerleider and M. Mishkin, "Two cortical visual systems. In: Ingle D j, Goodale M A, Mansfield R J W, eds. Analysis of Visual Behavior. Cambridge, MA: MIT Press, **(1982)**, pp. 549-586.
[9] M. Goodale and A. D. Milner, "Separate visual pathways for perception and action", Trends in neuroscience, vol. 15, **(1992)**, pp. 20-25.
[10] A. D. Milner and M. A. Goodale, "The visual brain in action. USA: Oxford University Press, **(1995)**, pp. 1-10.
[11] F. Fang, H. Boyaci and D. Kersten, "Border Ownership Selectivity in Human Early Visual Cortex and its Modulation by Attention", The Journal of Neuroscience, vol. 29, no. 2, **(2009),** pp. 460-465.
[12] B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio, "Categorization by learning and combining object parts," in Advances in Neural Information Processing Systems, vol. 14, **(2002)**.
[13] B. Leung, "Component-based car detection in street scene images," Master's thesis, EECS, MIT, **(2004)**.
[14] M. Weber, W. Welling, and P. Perona, "Unsupervised learning of models of recognition," in Proc. of the European Conference on Computer Vision, vol. 2, **(2000),** pp. 1001–108.
[15] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, **(2003),** pp. 264–271.
[16] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in Proc. IEEE Conf. on Computer Visionand Pattern Recognition, Workshop on Generative-Model Based Vision, **(2004).**
[17] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, **(2005)**.

[18] B. Leibe and B. Schiele, "Interleaved object categorization and segmentation," in BMVC, Norwich, UK, **(2003)** September, pp. 759–768.
[19] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, **(2001),** pp. 349–361.
[20] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in SLCP '04 Workshop on Statistical Learning in Computer Vision, **(2004)**.
[21] S. Ullman, M. Vidal-Naquet, and E. Sali, "Visual features of intermediate complexity and their use in classification," Nat. Neurosci., vol. 5, no. 7, **(2002),** pp. 682–687.
[22] I. A. Rybak, V. I. Gusakova and A. V. Golovan, "A model of attention-guided visual perception and recognition. Vision Research, vol. 38, **(1998),** pp. 2387-2400.
[23] D. G. Lowe, "Object recognition from local scale-invariant features," in Proc. of the International Conference on Computer Vision, **(1999),** pp. 1150–1157.
[24] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," Nat. Neurosci., vol. 2, no. 11, **(1999),** pp. 1019–25.
[25] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, "A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex," MIT, Cambridge, MA, AI Memo 2005-036 / CBCL Memo 259, **(2005)**.
[26] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex, J. Neurophys, vol. 58, **(1987),** pp. 1233–1258.
[27] T. Serre and M. Riesenhunber, "Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex," MIT, Cambridge, MA, CBCL Paper 239/ AI Memo 2004-2017, **(2004)**.

Authors

Zuojin Li is an Associate Professorsenior lecturer at the College of Electrical and Information Engineering, Chongqing University of Science and Technology in China. He received his PhD from the Chongqing University in China. Currently, He is working as a Post Doctor at computing department of UNITEC, New Zealand. His research interests cover machine vision, image processing, pattern recognition, intelligence system, multi-sensor data fusion.

Liukui Chen is an Associate Professorsenior lecturer at the College of Electrical and Information Engineering, Chongqing University of Science and Technology in China. He received his PhD from the Wuhan University in China. Currently. His research interests cover machine vision, image processing, pattern recognition, intelligence system, Intelligent Systems. He is the corresponding author of this paper.

Chen Gui is a senior lecturer at the College of Electrical and Information Engineering, Chongqing University of Science and Technology in China. He received his MSc. from the Chongqing University in China. Currently, He is working towards his Ph.D. in Computer Science from Aberystwyth University and his subject is the detection and acquisition of the science targets for planetary exploration. His research areas have focused on computer vision and machine learning technologies for autonomous science.