SCIENCE CHINA
Information Sciences

• LETTER •

December 2024, Vol. 67, Iss. 12, 229102:1–229102:2
https://doi.org/10.1007/s11432-022-4237-4

DcnnGrasp: towards accurate grasp pattern recognition with adaptive regularizer learning

Xiaoqin ZHANG[1], Ziwei HUANG[1], Jingjing ZHENG[2*],
Shuo WANG[2] & Xianta JIANG[2]

[1]*College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou* 325035*, China;*
[2]*Department of Computer Science, Memorial University of Newfoundland, St. John's* 36545*, Canada*

Received 11 December 2022/Revised 7 February 2024/Accepted 19 November 2024/Published online 13 December 2024

Vision-based grasp pattern recognition[1)] identifies the specific hand pose (fingers and palm configuration) for the object to be grasped based on visual information of household objects, which benefits tasks such as upper limb prosthesis control and gesture-based human-robot interaction. Various methods, including convolutional neural networks (CNN), have been proposed for grasp pattern recognition, leading to advancements in fields such as robot imitation, human-robot interaction, prosthetic design, and robot control. A pioneering CNN-based method called CnnGrasp was introduced, which inspired the development of subsequent CNN-based grasp pattern recognition methods [1]. However, traditional grasp pattern recognition methods often encounter challenges when tested with objects, and their views were not encountered by the algorithms during model training, which leads to degraded performance. In addition, compared with object categories labeling, grasp type labeling is much more complex and takes a lot of human resources, considering the varying hand-controlling requirements and gesture annotations of different applications and tasks. Motivated by the fact that object category usually contains distinctive information related to specific grasp types, including 3D structure, size, and even the material of the objects, this paper presents a novel approach that harnesses object category information to enhance the performance of grasp pattern recognition. Specifically, we propose a novel approach, namely the dual-branch CNN for grasp pattern recognition (DcnnGrasp), where object category classification is employed as an auxiliary task to enhance the effectiveness of grasp pattern recognition. To facilitate collaborative learning between the two tasks, we propose a new loss function called joint cross-entropy with adaptive regularizer (JCEAR), derived from maximizing a posterior. By exploiting the correlation between object category and grasp type, our method showed superior performance, strong generalizability and robustness compared to the state-of-the-art algorithms, particularly in scenarios involving unseen objects and datasets with limited grasp labeling.

DcnnGrasp for grasp pattern recognition. This study introduces a dual-branch network structure (DcnnGrasp) for the joint learning of object category classification and grasp pattern recognition. As illustrated in Figure 1, DcnnGrasp consists of two branches, each taking a household object image $\boldsymbol{x}$ as input. Each branch includes a feature extractor and a classifier. The features extracted from the object category feature extractor ($\boldsymbol{I}_{\text{category}}$) and the grasp feature extractor ($\boldsymbol{I}_{\text{grasp}}$) are integrated to enhance the performance of grasp pattern recognition. The object category feature extractor consists of convolutional layers (DenseNet), followed by m_{cf} fully connected layers.

Additionally, CnnGrasp is employed as the grasp feature extractor. The outputs of the category classifier and grasp classifier are defined as $\boldsymbol{f}^{o}(\boldsymbol{x};\theta_1,\theta_2)=[f_1^{\text{o}}(\boldsymbol{x};\theta_1,\theta_2),\ldots,f_{l_o}^{\text{o}}(\boldsymbol{x};\theta_1,\theta_2)]^{\text{T}}\in\mathbb{R}^{l_o}$ and $\boldsymbol{f}^{\text{g}}(x;\theta_1,\theta_3,\theta_4)=[f_1^{\text{g}}(\boldsymbol{x};\theta_1,\theta_3,\theta_4),\ldots,f_{l_g}^{\text{g}}(\boldsymbol{x};\theta_1,\theta_3,\theta_4)]^{\text{T}}\in\mathbb{R}^{l_g}$, respectively, where θ_1, θ_2, θ_3, and θ_4 represent the learnable parameters in the object category feature extractor, the category classifier, the grasp type feature extractor, and the grasp classifier, respectively. Here, l_o and l_g represent the numbers of object classes and grasp classes, respectively.

JCEAR. Since joint cross-entropy (JCE) simply combines the cross-entropy functions of the two tasks without capturing the relationship between object category classification and grasp pattern recognition, we propose JCEAR from a Bayesian perspective. This approach aims to enhance the collaborative learning of the two tasks and achieve improved performance.

Let one-shot vectors $\boldsymbol{c}^{\text{g}}=[c_1^{\text{g}},\ldots,c_{l_g}^{\text{g}}]^{\text{T}}$ and $\boldsymbol{c}^{\text{o}}=[c_1^{\text{o}},\ldots,c_{l_o}^{\text{o}}]^{\text{T}}$ be the true labels for grasp pattern recognition and object category classification, respectively. Our goal is to learn the parameters $\{\theta_k\}_{k=1}^4$ in $f_i^{\text{g}}(\boldsymbol{x};\theta_1,\theta_3,\theta_4)=p(c_i^{\text{g}}=1|\theta_1,\theta_3,\theta_4,\boldsymbol{x})$ and $f_j^{\text{o}}(\boldsymbol{x};\theta_1,\theta_2)=p(c_j^{\text{o}}=1|\theta_1,\theta_2,\boldsymbol{x})$ that maximize $p(\{\theta_k\}_{k=1}^4|\boldsymbol{c}^{\text{g}},\boldsymbol{c}^{\text{o}},\boldsymbol{x})$. Considering that $\boldsymbol{c}^{\text{g}}$ and $\boldsymbol{c}^{\text{o}}$ follow multinomial distributions, we define the likelihood function as $p(\boldsymbol{c}^{\text{g}},\boldsymbol{c}^{\text{o}}|\{\theta\}_{k=1}^4,\boldsymbol{x})=\prod_{i=1}^{l_{\text{o}}}f_i^{\text{o}}(\boldsymbol{x};\theta_1,\theta_2)^{c_i^{\text{o}}}\prod_{j=1}^{l_{\text{g}}}f_j^{\text{g}}(\boldsymbol{x};\theta_1,\theta_3,\theta_4)^{c_j^{\text{g}}}$. Introducing a matrix $\boldsymbol{W}\in\mathbb{R}^{l_{\text{o}}\times l_{\text{g}}}$, each element of the matrix $\boldsymbol{W}$ is de-

[*] Corresponding author (email: jjzheng233@gmail.com)

[1)] The grasp-type recognition refers to the ability of a system to identify the type of grasp a human or robot is using based on observed data. In contrast, the grasp-type detection task focuses more on perceiving a grasp action that has occurred or is occurring.

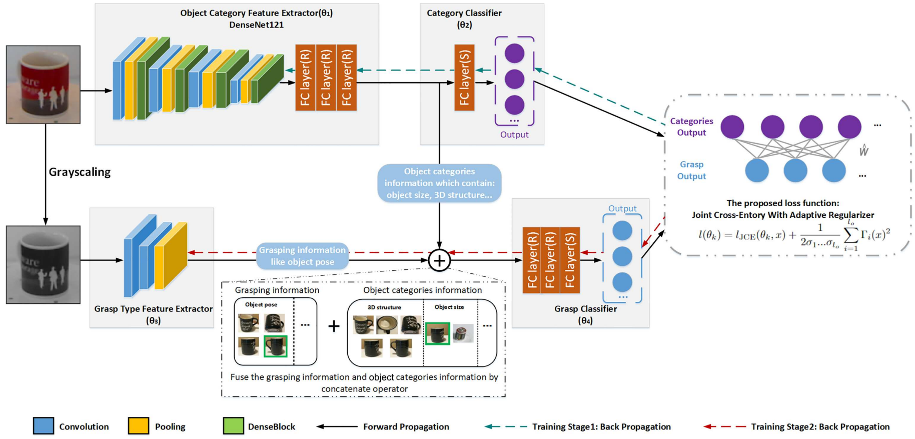
Object Category Feature Extractor(θ₁)
DenseNet121
FC layer(R)
FC layer(R)
FC layer(R)
Category Classifier
(θ₂)
FC layer(S)
Output
Categories Output
Grasp Output
Ŵ
The proposed loss function:
Joint Cross-Entory With Adaptive Regularizer
$l(\theta_k) = l_{\mathrm{JCE}}(\theta_k, x) + \frac{1}{2\sigma_1 \dots \sigma_{l_o}} \sum_{i=1}^{l_o} \Gamma_i(x)^2$
Grayscaling
Object categories information which contain: object size, 3D structure...
Grasping information like object pose
Grasp Type Feature Extractor
(θ₃)
Grasping information
Object categories information
Object pose
3D structure
Object size
Fuse the grasping information and object categories information by concatenate operator
FC layer(R)
FC layer(R)
FC layer(S)
Output
Grasp Classifier
(θ₄)
Convolution
Pooling
DenseBlock
Forward Propagation
Training Stage1: Back Propagation
Training Stage2: Back Propagation

Figure 1 (Color online) Architecture of DcnnGrasp and the training strategy based on JCEAR, where the parameters in the two branches are trained by Backpropagation stage 1 and stage 2, respectively.

fined as $W_{i,j} = p(c_i^{\mathrm{o}} = 1 | c_j^{\mathrm{g}} = 1, \{\theta_k\}_{k=1}^4, \boldsymbol{x})$, then we have $\boldsymbol{W} \boldsymbol{f}^{\mathrm{g}}(\boldsymbol{x}; \theta_1, \theta_3, \theta_4) = \boldsymbol{f}^{\mathrm{o}}(\boldsymbol{x}; \theta_1, \theta_2)$.

Calculating the distribution of $\boldsymbol{c}^o$ and $\boldsymbol{c}^g$ over the training set, we can estimate $W_{i,j}$ by $\hat{W}_{i,j} = \frac{N_{i,j}}{\sum_{q=1}^{l_g} N_{i,q}}$, where $N_{i,j}$ represents the number of the image samples belonging to the i-th class object and the j-th class grasp. Assuming that $\boldsymbol{f}^{\mathrm{o}}(\boldsymbol{x}; \theta_1, \theta_2) - \hat{\boldsymbol{W}} \boldsymbol{f}^{\mathrm{g}}(\boldsymbol{x}; \theta_1, \theta_3, \theta_4) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma})$, we define a prior as $p(\{\theta_k\}_{k=1}^4 | \boldsymbol{x}) = \frac{1}{(2\pi)^{l_o/2} \sigma_1 \cdots \sigma_{l_o}} \exp(\sum_{i=1}^{l_o} - \frac{\Gamma_i(\boldsymbol{x})^2}{2\sigma_i^2})$, where $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \dots, \sigma_{l_o}]$, $\Gamma_i(\boldsymbol{x}) = \boldsymbol{f}_i^o(\boldsymbol{x}; \theta_1; \theta_3; \theta_4) - \hat{\boldsymbol{W}}_{i,:} \boldsymbol{f}^g(\boldsymbol{x}; \theta_1; \theta_2)$ and $\hat{\boldsymbol{W}}_{i,:}$ is i-th row vector of $\hat{\boldsymbol{W}}$. Then, from the Bayesian perspective, $\hat{\theta}_k (k = 1, 2, 3, 4)$ can be obtained by

$$\arg\max_{\theta_k (k=1,2,3,4)} \ln(p(\{\theta_k\}_{k=1}^4 | \boldsymbol{c}^{\mathrm{g}}, \boldsymbol{c}^{\mathrm{o}}, \boldsymbol{x}))$$

$$= \arg\min_{\theta_k (k=1,2,3,4)} l_{\mathrm{JCE}}(\boldsymbol{x}, \{\theta_k\}_{k=1}^4) + \sum_{i=1}^{l_o} \frac{1}{2\sigma_i^2} \Gamma_i(\boldsymbol{x})^2,$$

where $l_{\mathrm{JCE}}(\boldsymbol{x}, \{\theta_k\}_{k=1}^4) = -\sum_{i=1}^{l_{\mathrm{o}}} c_i^{\mathrm{o}} \ln(f_i^{\mathrm{o}}(\boldsymbol{x}; \theta_1, \theta_2)) - \sum_{j=1}^{l_{\mathrm{g}}} c_j^{\mathrm{g}} \ln(f_j^{\mathrm{g}}(\boldsymbol{x}; \theta_1, \theta_3, \theta_4))$. It drives our loss function (JCEAR): $l(\boldsymbol{x}, \{\theta_k\}_{k=1}^4) = l_{\mathrm{JCE}}(\boldsymbol{x}, \{\theta_k\}_{k=1}^4) + \sum_{i=1}^{l_o} \frac{1}{2\sigma_i^2} \Gamma_i(\boldsymbol{x})^2$. The whole JCEAR-based training strategy for DcnnGrasp is illustrated in Figure 1 and detailed in Appendixes C and D.

Experiments. Experiments have been conducted on five publicly available household object datasets to verify the effectiveness of the proposed method, DcnnGrasp, in grasp classification. Detailed experimental results and discussions can be found in Appendix E. In the within-whole dataset cross-validation (WWC) scenario, DcnnGrasp consistently achieved the best results on all datasets, with all evaluation metric values exceeding 98%. For the unseen problem, our method achieved a global accuracy (GA) of about 94% and 99% on the RGB-D object dataset [2] and Hit-GPRec dataset [3], respectively, outperforming the second-best by nearly 15% in GA on the RGB-D object dataset. It demonstrates its potential for strong robustness in real-life applications. Additionally, even when there is only one object with gesture labels per object category in the training process, DcnnGrasp achieves GA values of over 90% (see Table E6 in Appendix E). It indicates that our proposed method can partially solve the challenge of difficult gesture labeling in grasp pattern recognition. Ablation studies revealed substantial improvements resulting from the proposed training strategy based on JCEAR. Notably, on the RGB-D object dataset and Hit-GPRec dataset, the increase in GA is at least 13.5%. These findings provide strong evidence supporting the effectiveness of the proposed strategies in enhancing grasp pattern recognition.

Conclusion. To address the challenge of handling unseen objects in grasp classification, we propose a novel dual-branch CNN (DcnnGrasp) and a new training strategy based on JCEAR. The simulated experimental results demonstrate the significantly superior performance of the proposed method to other state-of-the-art methods in grasp classification. In addition to its superior generalizability to unseen objects, our method also exhibits stronger robustness in handling 3D information of objects and datasets with only a few grasp labeling results. Therefore, introducing the object category information by combining dual-branch networks and the JCEAR training strategy holds great significance and application prospects in the field of grasp pattern recognition.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. U2033210) and Zhejiang Provincial Natural Science Foundation (Grant No. LDT23F02024F02).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Ghazaei G, Alameer A, et al. Deep learning-based artificial vision for grasp classification in myoelectric hands. J Neural Eng, 2017, 14: 036025

2 Lai K, Bo L F, Ren X F, et al. A large-scale hierarchical multi-view RGB-D object dataset. In: Proceedings of IEEE International Conference on Robotics and Automation, 2011. 1817–1824

3 Shi C Y, Yang D P, Zhao J D, et al. Computer vision-based grasp pattern recognition with application to myoelectric control of dexterous hand prosthesis. IEEE Trans Neural Syst Rehabil Eng, 2020, 28: 2090–2099