Robotics Overview

Peizhen Li[1]
https://lipzh5.github.io/ResearchProgress/

***Abstract*— A brief summary on top-conferences(ICRA, IROS, RSS, CoRL), top transactions/journals such as International Journal of Robotics Research (IJRR), IEEE Transactions on Robotics (TRO) and hot research topics.**

I. ICRA

Page limit: 6 + any number of pages for the bibliography/references (two-column format).

A. Themes:

- Soft Robot Applications
- Design of Mechanisms
- Planning
- Reinforcement Learning
- Marine and Field Robotics.
- Modeling, Control, and Learning for Soft Robots
- Compliant Mechanisms
- Path Planning and Collision Avoidance
- Deep Learning and Neural Networks in Robotics
- Manipulation and Grasping
- Human Centred and Inspired Robotics
- Deep Learning for Visual Perception
- Human-Robot Interation/Collaboration
- Computer Vision and Visual Servoing
- Optimal Control and Object Detection

B. Selected papers in 2023

- Code as policies: Language model programs for embodied control [1].

II. IROS

Page limit: 6, two-column format (up to two extra pages, $205 USD charge per extra page).

A. Themes and Selected papers

1) Cognitive robotics:

- (Winner) Gesture2Vec: Clustering Gestures using Representation Learning Methods for Co-speech Gesture Generation.
- Learning on the Job: Long-Term Behavioural Adaptation in Human-Robot Interactions.
- Intuitive & Efficient Human-robot Collaboration via Real-time Approximate Bayesian Inference.

2) Robot Mechanisms and Design:

- (Winner) Aerial Grasping and the Velocity Sufficiency Region.
- 1-degree-of-freedom robotic gripper with infinite self-twist function.

3) Entertainment and Amusement:

- (Winner) Robot Learning to Paint from Demonstrations.
- Robot Dance Generation with Music Based Trajectory Optimization.

4) Mobile Manipulation:

- (Winner) Robot Learning of Mobile Manipulation with Reachability Behavior Priors.
- Mobile Manipulation Leveraging Multiple Views.

III. RSS

Page limit: no limit, typically 8. Single-track, all aspects of robotics including scientific foundations, mechanisms, algorithms, applications, and analysis of robotic systems.

A. Paper Sessions:

- Human-Centered Robotics
- Manipulation from Demonstrations and Teleoperation
- Self-supervision and RL for Manipulation
- Large Data and Vision-Language Models for Robotics
- Simulation and Sim2Real
- Grasping and Manipulation
- Mobile Manipulation and Locomotion
- Robot Planning
- Robot State Estimation
- Robot Perception
- Control & Dynamics
- Robot Mechanisms & Control
- Autonomous Vehicles & Field Robotics
- Multi-Robot and Aerial Systems

IV. CoRL

Page limit: 8 pages + n pages for references.

A. Research areas

- Learning representations for robotic perception and control.
- Learning robot foundation models or general-purpose knowledge systems for robotics.
- Imitation learning for robotics, e.g. by behavioral cloning and/or inverse reinforcement learning.
- Reinforcement learning for control of physical robots.
- Model-based and model-free learning for robotic control and decision-making.
- Combination of learning- and planning-based approaches in robotics.
- Probabilistic learning and representation of uncertainty in robotics.
- Automatic robotic data generation for learning methods in robotics.

[1]Peizhen Li is with Faculty of Science and Engineering, School of Computing, Macquarie University, Macquarie Park NSW 2113 peizhen.li1@students.mq.edu.au

- Learning for Robot Task and Motion Planning.
- Learning for multimodal robot perception, sensor fusion, and robot vision.
- Learning for human-robot interaction and robot instruction by natural language, gestures as well as alternative devices.
- Learning for hardware design and optimization.
- Applications of robot learning in robot manipulation, navigation, locomotion, driving, flight, and other areas of robotics.
- Robot systems, hardware, and sensors for learning and data-driven approaches.

B. Selected papers

- Do as i can, not as i say: Grounding language in robotic affordances [2].
- Training Robots to Evaluate Robots: Example-Based Interactive Reward Functions for Policy Learning [3].
- BC-Z:Zero-Shot Task Generalization with Robotic Imitation Learning [4]

V. Popular Topics appeared on top trans/journals

A. IJRR

Dielectric Elastomer, Reinforcement Learning, Simultaneous Localization And Mapping,
Grasping, Multi Agent Systems, Biped Robot, Deep Learning, Adaptive Control, Model Checking, Teleoperation.

1) Selected papers:

- how to train your robot with deep reinforcement learning lessons we have learned [5].
- Human motion trajectory prediction: A survey [6].

B. TRO

Multi Agent Systems, Biped Robot, Simultaneous Localization And Mapping, Dielectric Elastomer, Grasping, Parallel Manipulator, Teleoperation, Reinforcement Learning, Myxococcus Xanthus, Adaptive Control.

C. Annual Review of Control, Robotics, and Autonomous Systems

Multi Agent Systems, Myxococcus Xanthus, Biped Robot, Teleoperation, Reinforcement Learning, Linear Matrix Inequalities, Complex Networks, Simultaneous Localization And Mapping, Human-robot Interaction, Dielectric Elastomer

D. Science Robotics

Dielectric Elastomer, Myxococcus Xanthus, Human-robot Interaction, Flapping Wing, Stretchable Electronics, Biped Robot.

VI. Research Direction

Robot Learning would be the focus. Specifically, aim at expanding robots' perception and physical interaction capabilities. Possible directions could be[1]:

- Multi-Modal Perception: Harnessing vision, touch, audio, and language for fine-grained and effective manipulation.
- Embodied Intelligence: Focusing on long-horizon planning, generalization to diverse environments, and sim-to-real transfer [7].
- Intuitive Physics: Learning structured world models for robotic manipulation of objects with diverse physical properties.

Appendix

A. Selected Notes

1) SayCan [2]:

- LLMs combined with Value Functions (for task affordances): LLMs help robots understand the high-level instructions and iteratively select useful and practical skills (low-level commands) until the task is finished.
- For example, given the task " I pilled my coke, can you bring me something to clean it up?", SayCan successfully planned and executed the following steps: 1. Find a sponge 2. Pick up the sponge 3. Bring it to you 4. Done

2) Robotics Transformer 1 [8]:

- Developed by researchers at Robotics at Google and Everyday Robots, 2022
- Transformer-based model,build upon a FiLM-conditioned EfficientNet, a TokenLearner, and a Transformer
- Trained with imitation learning with inputs of natural language tasks and images and output robot actions

3) Robotics Transformer 2 [9]:

- Builds upon VLMs that take one or more images as input, and produces a sequences of tokens representing natural language text. In order to control a robot, RT-2 represents robotic actions as tokens in the model's output - similar to language tokens (output action tokens to control a robot).
- Combine robotic control with chain-of-thought reasoning to enable leaning long-horizon planning and low-level skills within a single model.
- **Difference between SayCan and RT-2:** SayCan **can not see** the world and rely entirely on language while RT-2 can plan from both image and text commands.

Related language models in SayCan, RT-1, and RT-2: PaLI and PaLM [10], [11]

[1] https://yunzhuli.github.io/

4) Chain-of-Thought Prompting [12]:

- A simple mechanism for eliciting multi-step reasoning behavior in large language models. Motivated by **using intermediate steps to solve reasoning problems** and **few-shot prompting**.
- Does not positively impact performance for small models, and only yields performance gains when used with models of $\backsim$ 100B parameters.
- Has larger performance gains for more-complicated problems.

5) Long-horizon Planning [13]:

- Plans in the space of object subgoals, i.e., more **abstract** space of key object configurations, an idea well studied in Task-and-Motion planning (TAMP) [14].
- For rigid bodies, this abstraction can be realized using low-level manipulation skills that maintain **sticking contact** with the object and **represent subgoals as 3D transformations**.
- How to generalize to unseen objects? subgoal abstraction and representation.

6) Sim-to-Real Transfer [7]:

- A method to bridge the “reality gap”.
- Developing policies that are capable of adapting to very different dynamics by randomizing the dynamics of the simulator during training.

7) Vision Language Model [15]:

- **CLIP: Contrastive Language Image Pre-training**. Learning Transferable Visual Models From Natural Language Supervision.

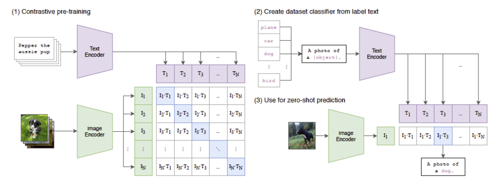

Fig. 1. summary of CLIP.

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Fig. 2. Numpy-like pseudocode for the core of an implementation of CLIP.

8) *Deep Reinforcement Learning:*

- BC-Z:Zero-Shot Task Generalization with Robotic Immitation Learning [4]
  - Study the problem of enabling a vision-based robotic manipulation system to generalize to novel tasks.
  - Approach the challenge from an **immitation learning perspective**.
  - Aiming to study how **scaling and broadening the data collected** can facilitate such gteneralization.
  - Policy training: given a fixed embedding, train $\pi(a|s,z)$ via Huber loss on XYZ and axis-angle predictions, and log loss for gripper angle.

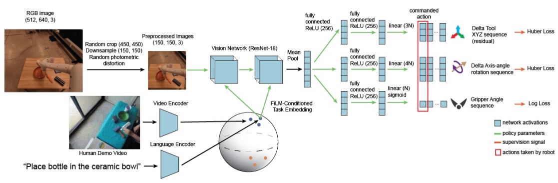

Fig. 3. A monocular RGB image from the head-mounted camera is passed through a ResNet 18 encoder, then through a two-layer MLP to predict each action modality (delta XYZ, delta axis-angle, and gripper angle). FiLM layers condition the architecture on a task embedding z computed from language w_l or video w_h

Algorithm 1: Pseudocode for training the video encoder

Input: Task commands $\mathcal{W}$, per-task robot dataset $\mathcal{D}_e^i$, per-task human video data $\mathcal{D}_h^i$, language encoder $q(\cdot|w_\ell^i)$, video encoder $q(\cdot|w_h)$

```
while not done training do
    Sample a batch of tasks i, with replacement.
    for each task i ∈ batch do
        Sample human video w_h ∈ D_h^i
        Sample robot demo {(s_t, a_t)}_{t=1}^T ∈ D_e^i
        Retrieve language command w_ℓ^i
        z_h^i ~ q(·|w_h)    // embed human video
        z_e^i ~ q(·|{s_t}_{t=1}^T)    // embed robot video
        z_ℓ^i ~ q(·|w_ℓ^i)    // get language vector
        Sample t ∈ 1, ··· , T
        Compute action π(â|s_t, z_h^i)
        BC-loss ← 100 · Huber(xyz) + 10 · Huber(angle) + 0.5 · LogLoss(gripper)
        Minimize L ← BC-loss + D_cos(z_h^i, z_ℓ^i) + D_cos(z_e^i, z_ℓ^i)
    end
end
```

Fig. 4. BC-Z: pseudocode for training encoder.

- MT-Opt-Continuous Multi-Task Robotic Reinforcement Learning at Scale [17]

  a) *Takeaway:*

- UCL Course on RL
- Andrej Karpathy's blog on RL
- Reinforcement Learning 101

9) *Imitation Learning:*

10) *Large Language Model [18]:*

- How to achieve the "meta-learning" or "in-context learning" ability?
- In-context learning approach
- GPT-2

  a) *Pathways Language Models [19]:*

- Concepts of zero-shot and few-shot prompting.
- GPT3-few shot learner for language model.
- Pathways, a next-generation AI architecture.

11) Visual Reasoning with a General Conditional Layer [16]:

- **FiLM:** Feature-wise Linear Modulation.
  - Influence neural network computation via a simple, feature-wise affine transformation based on conditioning information
  - Can be viewed as using one network to generate parameters of another network, making it a form of hypernetwork.

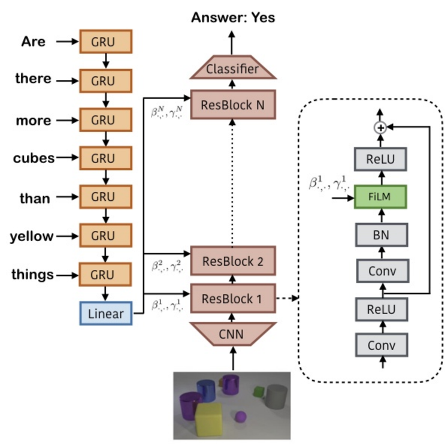

Fig. 5. The FiLM Generator (left), FiLM-ed network (mid), and residual block architecture (right).

12) TokenLearner [20]:

- **TokenLearner:** Adaptive Space-Time Tokenization for Videos
  - A novel visual representation learning that learns to mine important tokens in visual data.
  -

$$z_i = A_i(X_t) = \rho(X_t \odot A_{iw}) = \rho(X_t \odot \gamma(\alpha_i(X_t))) \quad (1)$$

where $\odot$ is the Hadmard product (i.e., element-wise multiplication), and $A_{iw} \in \mathbb{R}^{H \times W \times C}$ is an intermediate weight tensor computed with the function $\alpha_i(X_t)$ and the broadcasting function $\gamma(\cdot)$

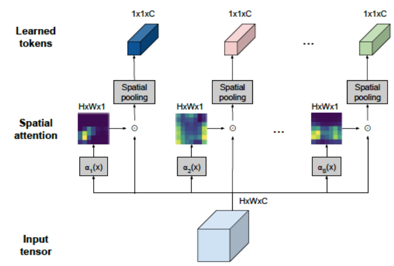

Fig. 6. Visual illustration of the TokenLearner module, applied to a single image frame..

ACKNOWLEDGMENT

References

[1] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9493–9500.

[2] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[3] K. Huang, E. S. Hu, and D. Jayaraman, "Training robots to evaluate robots: Example-based interactive reward functions for policy learning," *arXiv preprint arXiv:2212.08961*, 2022.

[4] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.

[5] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, "How to train your robot with deep reinforcement learning: lessons we have learned," *The International Journal of Robotics Research*, vol. 40, no. 4-5, pp. 698–721, 2021.

[6] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.

[7] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3803–3810.

[8] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[9] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[10] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv preprint arXiv:2209.06794*, 2022.

[11] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[13] A. Simeonov, Y. Du, B. Kim, F. Hogan, J. Tenenbaum, P. Agrawal, and A. Rodriguez, "A long horizon planning framework for manipulating rigid pointcloud objects," in *Conference on Robot Learning*. PMLR, 2021, pp. 1582–1601.

[14] L. P. Kaelbling and T. Lozano-Pérez, "Hierarchical task and motion planning in the now," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 1470–1477.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[16] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[17] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman, "Mt-opt: Continuous multi-task robotic reinforcement learning at scale," *arXiv preprint arXiv:2104.08212*, 2021.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[19] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*,

“Palm: Scaling language modeling with pathways,” *arXiv preprint arXiv:2204.02311*, 2022.

[20] M. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, “TokenLearner: Adaptive Space-Time Tokenization for Videos,” in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 12 786–12 797. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/6a30e32e56fce5cf381895dfe6ca7b6f-Paper.pdf