DexVIP: Learning Dexterous Grasping with Human Hand Pose Priors from Video

Priyanka Mandikal[1,2] and Kristen Grauman[1,2]
[1]UT Austin, [2]Facebook AI Research

Website: http://vision.cs.utexas.edu/projects/dexvip-dexterous-grasp-pose-prior

Motivation

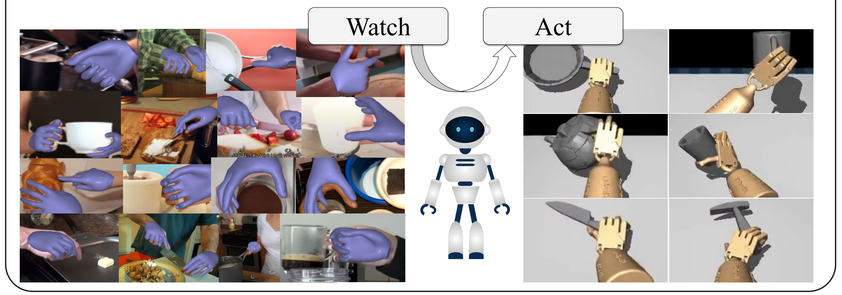
- Everyday objects are designed for human hands
- Dexterous robots are morphologically similar to human hands
  → we can derive vital cues for manipulation from video

Existing Methods

- Parallel jaw grippers to execute simpler 6-DOF grasps
  → simple pick-and-place tasks
- Dexterous grasps - tele-operation [1] / visual affordances [2]
  → costly supervision, poor scalability, unrealistic hand poses

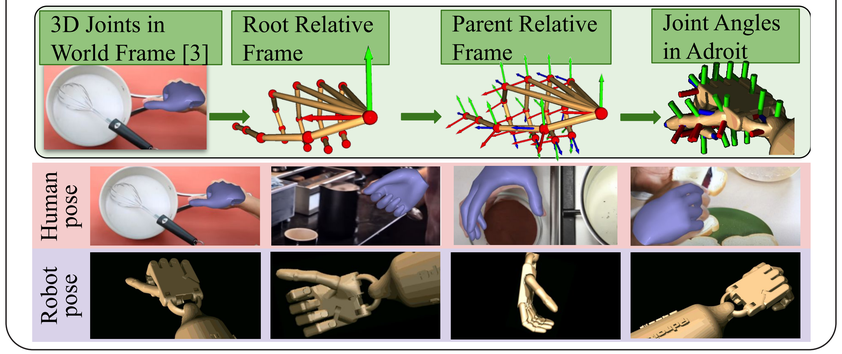
Our Idea

- How can robot grasping benefit from watching humans in action?
- Using hand poses extracted from in-the-wild YouTube video frames → train a dexterous robotic agent to grasp objects
- Focus not only on *where* but also *how* to grasp the object!



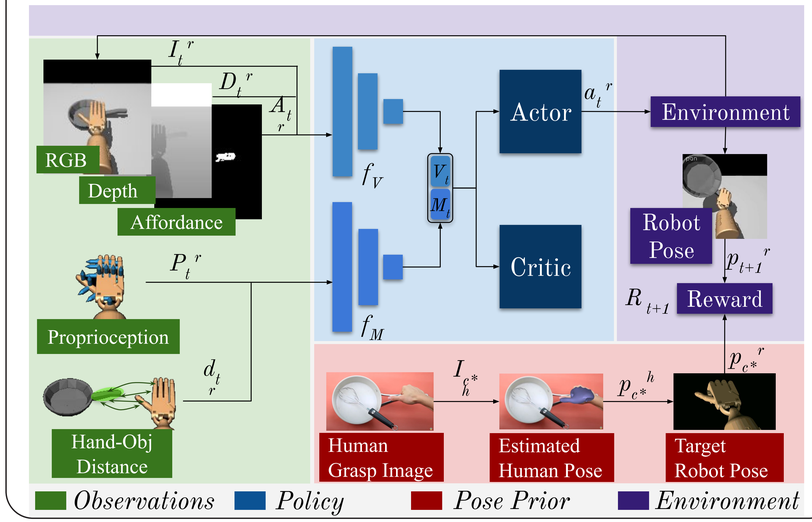
Pose Priors

Pose retargeting from human hand to robot hand

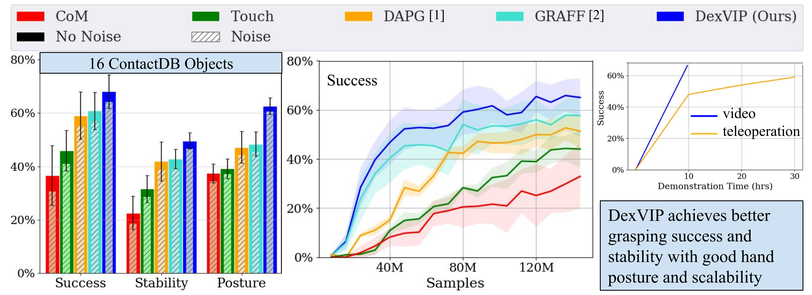


Overview

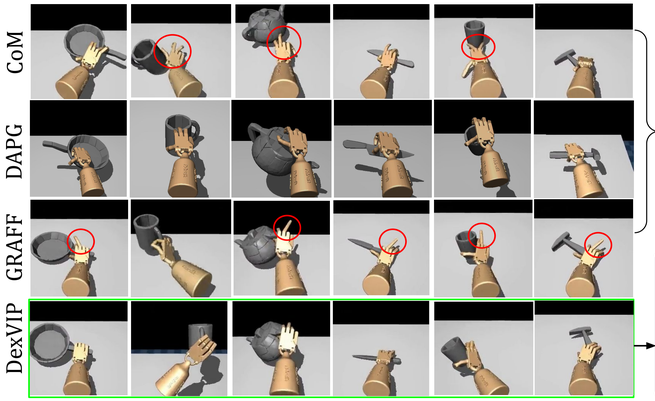
Visuo-motor policy rewards the agent for reaching target grasp pose



Results



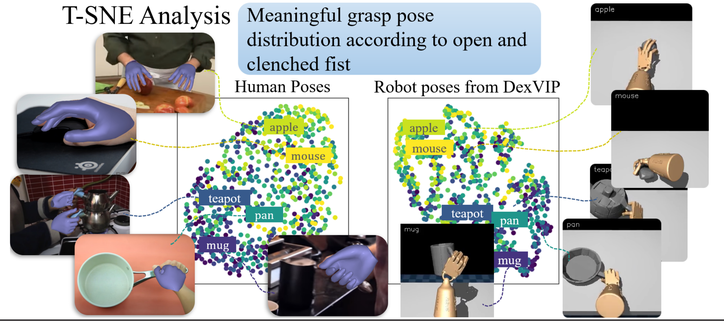
DexVIP achieves better grasping success and stability with good hand posture and scalability



- Unnatural poses lead to poor grasp stability
- Costly MoCap supervision for DAPG
- Pose priors from video enable learning good quality stable grasps

T-SNE Analysis

Meaningful grasp pose distribution according to open and clenched fist



Refs: [1] Learning complex dexterous manipulation with deep RL & demonstrations, RSS 2018 [2] GRAFF: Learning dexterous grasping using object-centric visual affordable, ICRA 2021 [3] FrankMocap: A monocular 3D whole-body pose estimation system via regression & integration, ICCV-W 2021