IITB-Monash Research Academy

An Indian-Australian research partnership

Indian Institute of Technology Bombay | MONASH University

| | |
|---|---|
| Project Title: | Joint 3D Reconstruction of Human and Objects in a Dynamic Scene using Monocular Video |
| ***Project Number*** | **IMURA0986** |
| ***Monash Main Supervisor*** *(Name, Email Id, Phone)* | Hamid Rezatofighi, |
| ***Monash Co-supervisor(s)*** *(Name, Email Id, Phone)* | Jianfei Cai, jianfei.cai@monash.edu, |
| ***Monash Head of Dept/Centre*** *(Name, Email)* | Jianfei Cai, jianfei.cai@monash.edu, |
| ***Monash Department:*** | Data Science & AI |
| ***Monash ADRT*** *(Name, Email)* | Timothy Scott |
| ***IITB Main Supervisor*** *(Name, Email Id, Phone)* | Biplab Banerjee, bbanerjee@iitb.ac.in |
| ***IITB Co-supervisor(s)*** *(Name, Email Id, Phone)* | Subhasis Chaudhuri, sc@ee.iitb.ac.in |
| ***IITB Head of Dept*** *(Name, Email, Phone)* | Kishore Chatterjee, kishore@ee.iitb.ac.in |
| ***IITB Department:*** | Department of Electrical Engineering |

| ***Research Clusters:*** | | ***Research Themes:*** | |
|---|---|---|---|
| ***Highlight which of the Academy's CLUSTERS this project will address?*** *(Please nominate JUST one. For more information, see www.iitbmonash.org)* | | ***Highlight which of the Academy's Theme(s) this project will address?*** *(Feel free to nominate more than one. For more information, see www.iitbmonash.org)* | |
| 1 | *Material Science/Engineering (including Nano, Metallurgy)* | 1 | ***Artificial Intelligence and Advanced Computational Modelling*** |
| 2 | *Energy, Green Chem, Chemistry, Catalysis, Reaction Eng* | 2 | *Circular Economy* |
| 3 | *Math, CFD, Modelling, Manufacturing* | 3 | *Clean Energy* |
| 4 | *CSE, IT, Optimisation, Data, Sensors, Systems, Signal Processing, Control* | 4 | *Health Sciences* |
| 5 | *Earth Sciences and Civil Engineering (Geo, Water, Climate)* | 5 | *Smart Materials* |
| 6 | *Bio, Stem Cells, Bio Chem, Pharma, Food* | 6 | *Sustainable Societies* |
| 7 | *Semi-Conductors, Optics, Photonics, Networks, Telecomm, Power Eng* | | |
| 8 | *HSS, Design, Management* | | |

| | | | |
|---|---|---|---|
| | | | |

## *The research problem*

### *Joint 3D Reconstruction of Human and Objects in a Dynamic Scene using Monocular Video*

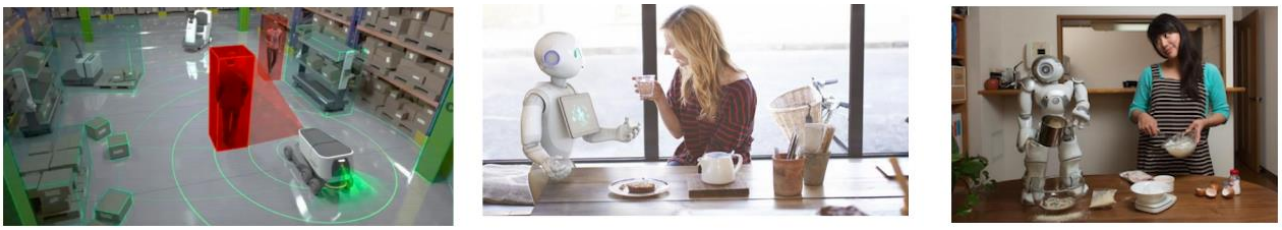

*Figure 1: (a) Robot perception during navigation (b) Robot's interaction with scene elements (e.g Social Robots) (c) Training a robot (learning from the interactions between scene elements)*

*3d dynamic scene reconstruction is an important area of work for robot perception and navigation, robot manipulation, and training a robot using reinforcement learning. To navigate or interact with the environment, the robot should have an accurate knowledge of the 3d models, which includes the shape, size, and exact location or layout of the scene elements (objects and humans). Joint 3d reconstruction of humans and objects in dynamic scenes is a challenging problem. A human can generally perceive its surrounding environment and correctly estimate object locations, shape, and size from some prior knowledge about the object, object-object, or human-object interactions. This knowledge may include information about which elements generally interact, the type of interactions (gravitational constraints or affordance), and relative positions of the elements during interactions. A human also perceives the relative distance between objects by reasoning the relative size of the elements. Hence, a 3d dynamic scene reconstruction system should consider all these attributes for an accurate and physically plausible reconstruction of scene elements.*

*Owing to the advancements in machine learning techniques and the availability of large-scale data, the recent developments [1,2] show impressive results on 3d scene reconstruction. However, mostly these methods focus on human and object reconstruction independently. These methods do not exploit the human-object interactions, although they can leverage important information for reconstructing individual scene elements. Some very recent methods [3, 4, 5] have considered joint 3d reconstruction of humans and objects in a scene. In the first work in this area [3], authors considered human-object interactions using some prior knowledge on HOI, physical plausibility of the object and human pose by imposing constraints on human-object/ object-object collision (except on container-like objects chair, desk), and support relations (between objects like monitor and table or table and floor, human and floor, etc.). This method only considers 3d skeleton and coarse level object reconstruction which limits its scope, as the exact shape or occupancy of objects or humans remains unknown. In [4], authors reconstruct human mesh and the objects in contact with a human. This paper considers the relative size of objects with respect to the human and depends on prior knowledge about the object sizes to finetune the object location and object layout. [5] is the first work that proposes a rich reconstruction of the*

*scene considering humans and most of the scene objects. This unsupervised method imposes 2d-3d consistency loss for optimizing each element independently. It also incorporates physical losses (human-object collision, human-object contact, ground plane supporting loss) to jointly optimize the pose and layout of the different scene elements. 2d-3d consistency loss can not resolve size and exact location ambiguities as many 3d predictions can project into the same 2d.*

*Overall, the state-of-art methods for 3d scene reconstruction only consider a prior knowledge of human-object interactions in a static scene with a single person environment, which may not be sufficient for a dynamic and multi-person environment. Context information about the activities going on in a scene can help to build more sophisticated physical constraints for better reconstruction. Also, current methods consider a limited number of objects and are highly dependent on the accuracy of 2D object detection algorithms.*

## *Project aims*

- ***Finding physical constraints for plausible 3d reconstruction of humans and objects in the dynamic scene****: Physical constraints like gravity (e.g., object rest on the floor), affordance (e.g., human can sit on the chair), the geometry of object (to handle object occlusions) should be considered for the plausible reconstruction of the scene elements. Some recent methods [12, 13] show that the learned human-scene interactions like surface contacts, penetration, affordance significantly improve the human pose estimation. However, we aim to leverage this information from images for joint reconstruction of humans and scenes. Constraints considering the relative size or relative distance between different scene elements (human and objects) can also give a better understanding of the size and shape of individual elements. Moreover, the surface contacts or interactions amongst the scene elements may change over time (e.g., a human moving object from one place to another, a person jumping on the floor). Hence, understanding the scene deformation or the temporal correlation across frames and spatial correlation amongst scene elements would lead to a better scene reconstruction.*

- ***Using multi-modal information for scene reconstruction****: In [2, 3, 4, 5], authors incorporate the object-object, object-scene, or human-object relationship implicitly. These methods often show object penetration or incorrect surface contacts (floating objects) in the final reconstruction. An additional mode of information along with the video, like video description or video captioning, can be used to have better context about the scene elements interactions. Dense video captioning datasets like MSR-VTT [8], charades [9], Activitynet caption [10], can be useful for this approach. "A person sits on a bed and puts a laptop into a bag" gives cues for exact positional relation between objects (person, bed, laptop, bag). This explicit object-scene relation can be used at the training time for self-supervision to learn the relative position of the objects or humans, which will eventually help to incorporate the physical constraints in a better way.*

- ***Making the 3d scene reconstruction robust for unseen objects****: For all the current methods, a major limitation is, these methods can detect a limited variety of objects or highly dependant on the 2D object detector, which is used to get an initial estimate of the scene objects location. In a dynamic scene, feature similarity between consecutive frames for the un-reconstructed regions can help recover the unknown objects. Zero-shot learning [11] for unseen object detection can be adopted in 3d scene reconstruction pipeline to adapt the unseen objects incrementally.*

- ***Building a socially feasible 3d scene reconstruction under multi-person environment****: Recent methods do not consider the multi-person environment in joint human-object reconstruction. The position or velocity of a person in a dynamic environment is often conditioned by the location of the other scene elements or surrounding people. In this project, we aim to exploit the social interactions (e.g., group of people sitting around a table), body-part interactions (e.g., hand-*

*shaking), known human body priors under multi-person environment for reasoning the physical constraints, person-person occlusion, or environmental occlusions for socially plausible 3d reconstruction.*

## *Expected outcomes*

*Physically plausible robust human and object mesh construction in a dynamic scene given a monocular video input.*

## *How will the project address the Goals of the above Themes?*

*This project will fit into the Advanced computational engineering, simulation and manufacturing theme. This project mainly focuses on robot perception. Robotics is one of the emerging technologies in AI with wide applications in healthcare, education, entertainment, manufacturing industries. The basic prerequisite for a robotic application is to provide a clear representation of the surrounding environment to the robot to make it able to perceive the environment better. This project focuses on creating the scene representation in 3D which gives a complete scene understanding i.e pose, shape and size of different scene elements (humans + objects) and Spatio-temporal interactions between them.*

## *Capabilities and Degrees Required*

*List the ideal set of capabilities that a student should have for this project. Feel free to be as specific or as general as you like. These capabilities will be input into the online application form and students who opt for this project will be required to show that they can demonstrate these capabilities.*

**Essential Skills:**

an Honour/Master degree in Computer Science, Mathematics, Physics, or Engineering.

A background in machine learning and/or computer vision.

Programming skills in a variety of coding languages (e.g., Python, Matlab, C/C++/C#) and proficient in one of the main deep learning libraries (e.g., Darknet, TensorFlow, PyTorch, Keras)

fluent communication skills in English.

**Desirable Skills**

Previous experience in 3D vision, e.g. 3D object reconstruction, human face or body reconstruction, or working with state-of-the-art computer vision methods in 3D vision

Showing the ability to publish a paper in top-tier (CORE A*/A) venues in computer

## *Reference:*

*[1] Nie Y, Han X, Guo S, Zheng Y, Chang J, Zhang JJ. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020.*

*[2] Zhang C, Cui Z, Zhang Y, Zeng B, Pollefeys M, Liu S. Holistic 3D Scene Understanding from a Single Image with Implicit Representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021.*

*[3] Chen Y, Huang S, Yuan T, Qi S, Zhu Y, Zhu SC. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2019.*

*[4] Zhang JY, Pepose S, Joo H, Ramanan D, Malik J, Kanazawa A. Perceiving 3d human-object spatial arrangements from a single image in the wild. In European Conference on Computer Vision (ECCV) 2020.*

*[5] Weng Z, Yeung S. Holistic 3D Human and Scene Mesh Estimation from Single View Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021.*

*[6] Sadeghian A, Kosaraju V, Sadeghian A, Hirose N, Rezatofighi H, Savarese S. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019.*
*[7] Chen DZ, Gholami A, Nießner M, Chang AX. Scan2Cap: Context-aware Dense Captioning in RGB-D Scans. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2019.*
*[8] J. Xu, T. Mei, T. Yao, and Y. Rui, MSR-VTT: A large video description dataset for bridging video and language, in 2016 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR) 2016.*
*[9] Sigurdsson GA, Varol G, Wang X, Farhadi A, Laptev I, Gupta A. Hollywood in homes: Crowdsourcing data collection for activity understanding. In European Conference on Computer Vision (ECCV) 2016.*
*[10] Krishna R, Hata K, Ren F, Fei-Fei L, Carlos Niebles J. Dense-captioning events in videos. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2017.*
*[11] Zhu P, Wang H, Saligrama V. Don't Even Look Once: Synthesizing Features for Zero-Shot Detection. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020.*
*[12] Hassan M, Choutas V, Tzionas D, Black MJ. Resolving 3D human pose ambiguities with 3D scene constraints. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2019.*
*[13] Hassan M, Ghosh P, Tesch J, Tzionas D, Black MJ. Populating 3D Scenes by Learning Human-Scene Interaction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021.*